Bayesian Modeling and Adaptive Monte Carlo with Geophysics Applications

by

Jianyu Wang

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Robert Wolpert, Co-Supervisor

_____
Scott Schmidler, Co-Supervisor

_____
Jim Berger

_____
Elaine Spiller

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2013

Abstract

Bayesian Modeling and Adaptive Monte Carlo with Geophysics Applications

by

Jianyu Wang

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Robert Wolpert, Co-Supervisor

_____
Scott Schmidler, Co-Supervisor

_____
Jim Berger

_____
Elaine Spiller

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University
2013

Abstract

The first part of the thesis focuses on the development of Bayesian modeling motivated by geophysics applications. In Chapter 2, we model the frequency of pyroclastic flows collected from the Soufriére Hills volcano. Multiple change points within the dataset reveal several limitations of existing methods in literature. We propose Bayesian hierarchical models (BBH) by introducing an extra level of hierarchy with hyper parameters, adding a penalty term to constrain close consecutive rates, and using a mixture prior distribution to more accurately match certain circumstances in reality. We end the chapter with a description of the prediction procedure, which is the biggest advantage of the BBH in comparison with other existing methods. In Chapter 3, we develop new statistical techniques to model and relate three complex processes and datasets: the process of extrusion of magma into the lava dome, the growth of the dome as measured by its height, and the rockfalls as an indication of the dome's instability. First, we study the dynamic Negative Binomial branching process and use it to model the rockfalls. Moreover, a generalized regression model is proposed to regress daily rockfall numbers on the extrusion rate and dome height. Furthermore, we solve an inverse problem from the regression model and predict extrusion rate based on rockfalls and dome height.

The other focus of the thesis is adaptive Markov chain Monte Carlo (MCMC) method. In Chapter 4, we improve upon the Wang-Landau (WL) algorithm. The WL algorithm is an adaptive sampling scheme that modifies the target distribution

to enable the chain to visit low-density regions of the state space. However, the approach relies heavily on a partition of the state space that is left to the user to specify. As a result, the implementation and the use of the algorithm are time-consuming and less automatic. We propose an automatic, adaptive partitioning scheme which continually refines the initial partition as needed during sampling. We show that this overcomes the limitations of the input user-specified partition, making the algorithm significantly more automatic and user-friendly while also making the performance dramatically more reliable and robust. In Chapter 5, we consider the convergence and autocorrelation aspects of MCMC. We propose an Exploration/Exploitation (XX) approach to constructing adaptive MCMC algorithms, which combines adaptation schemes of distinct types. The exploration piece uses adaptation strategies aiming at exploring new regions of the target distribution and thus improving the rate of convergence to equilibrium. The exploitation piece involves an adaptation component which decreases autocorrelation for sampling among regions already discovered. We demonstrate that the combined XX algorithm significantly outperforms either original algorithm on difficult multimodal sampling problems.

To my parents, my husband and my son.

Contents

List of Tables

List of Figures

Acknowledgements

First and foremost, I would like to express my gratitude and appreciation to my advisors Robert Wolpert, Jim Berger and Scott Schmidler, without whom this dissertation would simply be impossible. I thank Robert and Jim for introducing me to the research topic at the very early stage of my graduate study, for their patience and valuable guidance. I am especially grateful to them for all their understanding and support during the most difficult times in the past two years. I thank Scott for extending and broadening the course projects into interesting research topics that become the second part of this dissertation. I appreciate Scott's insights and rigorous attitude on improving the results.

I thank Susie Bayarri for her insightful comments on the first part of my dissertation. I feel sorry to learn about her health issue. I am grateful to her for serving my external committee member till the summer of 2013 and I sincerely wish her to get better and stay healthy. I thank Elaine Spiller for replacing Susie as my new external committee member and for her valuable suggestions and ideas on some topics of my dissertation.

I would like to thank all the faculty members and fellow graduate students for the great atmosphere during my graduate study. I thank Dalene Stangl for her help on my summer teaching. I thank Alan Gelfand and Mike West for their support on behalf of the department during the difficult times of my family. I thank David Banks and the department for hosting the Thanksgiving dinner and Halloween party. I feel

fortunate to know and make friends with many kind and smart people like Yingbo, Fangpo, Hongxia, Yajuan, Xiaojing, Lin, Hao, Kai, Francesca, Silvia and many more to list. I learn from them and enjoy their company.

This dissertation work was supported by the National Science Foundation grant DMS-00757549, the Statistical and Applied Mathematical Science Institute, and Duke University graduate school. I am also grateful to Duke University graduate school for the travel support I received to attend conferences and workshops.

Last, I would like to thank my family for their unconditional support and love. To my parents, Wei Wang and Xiufang Meng, for raising me and educating me. To my brother, Xiaodong Wang, for taking care of my parents while I am far away from home. I thank my parents and my parents-in-law for helping me take care of my son while I tried to finish the dissertation. I thank my husband, Bo Zhang, for believing in me, supporting me, and encouraging me all these years. Finally, special thanks to my wonderful son, William, for fighting so hard for life and being inspiringly brave and optimistic all the time.

1

Introduction

This thesis focuses primarily on Bayesian statistical methods. In particular, the research includes two independent pieces: Bayesian statistical modeling for non-stationary Markov processes and adaptive Markov chain Monte Carlo (MCMC) algorithms for efficient sampling. In Chapters 2 and 3, the focus of research is on Bayesian analysis of point processes in the geoscience application. Whereas the motivation and application of the methodology is associated with volcanic data, the theoretical framework of the problems is non-stationary Markov processes and the resulting methodology has broad applications. In Chapters 4 and 5, we develop an adaptive MCMC algorithm which has advantage in both exploration and exploitation of the state space.

1.1 Bayesian hierarchical modeling for change points in a point process

The point process data we analyze is a time series of pyroclastic flows from Soufrière Hills volcano. The scientific question is to determine the probability of the occurrence of certain catastrophic events at any time and any place on the island. We develop

statistical tools to assess and predict the risk associated with geophysical hazards such as volcanic pyroclastic flows. A particular goal is to study how these risks vary in space and time, and of how uncertain they are. The focus in the first part of my dissertation work is on Bayesian inference of stochastic processes with jumps.

A feature that makes the problem difficult to do well is that the process is very non-stationary—multiple change points in the process. We have been developing a change point model to reflect this non-stationarity, and have been using advanced statistical methods, Reversible jump Markov Chain Monte Carlo, to help support predictions that reflect all the uncertain aspects of the model and data. For practical purposes, We introduce a flexible penalized mixture prior distribution and subsequently apply Monte Carlo integration method to deal with difficulty in the calculation of normalizing constants.

We apply this Bayesian hierarchical modeling to investigate the changes in the eruption frequency of the volcano and predict the probability of future catastrophic events. The overall results of the real data show that the estimates coincide with significant geological changes of the volcano.

The methodology is developed in the context of a specific problem of pyroclastic flows, but it is applicable more broadly to problems in the analysis of other time-varying point data and quantification of other geohazards, risk, and etc.

1.2 Non-stationary Markov process with heavy-tailed distributions

Under the same context, We also analyze small, easy-to-detect volcanic events such as rockfalls. Furthermore, we discover the relationship between rockfalls and dome information, dome height and the extrusion rate (which cannot be directly observed). Hence, we are able to solve an inverse problem: To predict the dome's extrusion rate from observed rockfall numbers and dome height.

We propose to model the rockfall counts as a non-stationary correlated negative

binomial process for the following three reasons: (1) The distribution of the daily number of rockfalls has a much heavier tail than Poisson distribution. (2) Rockfall counts vary significantly at some time. (3) Rockfall counts for two consecutive days are usually highly correlated. Since the non-stationarity of the process can be reflected in different parameters of the marginal negative binomial distributions, we apply Bayes factor to compare three possible models. My dissertation research includes challenging problems involved in these models such as inference and simulation methods.

We develop a Bayesian generalized regression model for this non-stationary Markov process, based on the suggestion from the geologist in our research group that it is probable that the rockfall activities follow an underlying physical process induced by lava domes. Building a joint model for these processes is not only important for understanding the geophysical process, but also help develop methods to predict or estimate important geophysical quantities like the extrusion rate, key for predicting pyroclastic flows, from easy-to-measure features like dome height and rockfalls. Furthermore, considering an inverse problem, we establish a change point model to predict the extrusion rate based on rockfalls and dome height.

In application, our work supports important decision processes by developing new statistical techniques to model and relate three complex processes and data sets: the process of extrusion of magma into the lava dome that is the cause of pyroclastic flows, the growth of the dome as measured by its height, and the rockfalls which are small rock avalanches off the dome that are an indication of the dome's instability.

1.3 Hybrid strategies for adaptive Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC) methods are widely used in Bayesian statistics to sample from complex distributions. However, straightforward construction of MCMC chains such as Gibbs sampling or Metropolis-Hastings often requires sig-

nificant hand-tuning and mixes impractically slowly when the target distribution is high-dimensional, complicated, or multimodal. Many adaptive MCMC (AMCMC) schemes have been developed to address these issues. Recent theory suggests that adaptive schemes can be classified into two distinct types: those which aid exploration of the target distribution, and those which improve mixing among previously visited regions of the state space.

The contribution of my dissertation work to this area is a hybrid strategy—an Exploration/Exploitation (XX) approach—to constructing adaptive MCMC algorithms, which combines adaptation schemes of distinct types. One piece, the "exploration" piece, uses adaptation strategies aimed at exploring new regions of the target distribution and thus improving the rate of convergence to equilibrium. The other piece, the "exploitation" piece, involves an adaptation component which decreases autocorrelation for sampling among regions which are already discovered. This hybrid combination is relatively simple, yet provides the best of both worlds. As an example of this approach, we develop an XX algorithm that combines an adaptive Metropolized independence sampler (AMIS) as the exploitation component, with the generalized Wang-Landau (GWL) algorithm as the exploration component. We show that, for multimodal target distributions, both WL and AMIS algorithms require general purpose modifications.

WL algorithm partitions the state-space into subsets according to energy (log-density), and adaptively estimates marginal probabilities of sets, in order to reweight the target distribution on each component to achieve uniform sampling across sets. However, the performance relies heavily on the predefined energy partition of the state space which is left to the user to specify and may fail even for simple low-dimensional bimodal distributions. We have shown that this is due to fundamental restrictions on convergence imposed by the width of energy bins, and maximum energy height, specified by the initial partition. To overcome the limitations, we

develop an automatic, adaptive partitioning scheme which continually refines the initial partition as needed during sampling, making the algorithm significantly more automatic and user-friendly while also making the performance dramatically more reliable and robust for exactly those multimodal problems which WL/GWL sampling is designed to address.

Another type of adaptive MCMC approaches have been developed in which the transition kernel of the chains is sequentially modified over time based on the current sample history. Although they improve autocorrelation for chains, it is shown that they fail to improve convergence rates for multimodal distributions. AMIS belongs to this type. In combining the AMIS algorithm with the WL algorithm to form the XX algorithm, it reveals a previously unobserved breakdown of the AMIS algorithm on multimodal distributions. This weakness lies in the stochastic approximation (SA) formulation of AMIS. In particular, the sequential updating scheme of the AMIS mixture proposal does not accommodate very well large changes in the target distribution observed late in the sampling, which limits the ability to take full advantage of the information provided by the exploration chain in the context of the XX algorithm. In addition, the user-defined value of M, the number of mixture components in the proposal distribution, is arbitrary and it is desirable to ensure the algorithm is not handicapped by choosing M too low.

Our modifications of the algorithm allow user to address this easily, by starting with M small and adding one or more additional components whenever a new mode or region is identified. Note that we need not determine the optimal number of components, a notoriously difficult problem, but simply add components as needed to ensure M is sufficiently large. In addition, XX algorithm exhibits the complementary strengths of both methods: the ability of the Wang-Landau algorithm to cross arbitrary energy barriers, and the ability of the AMIS algorithm to dramatically reduce autocorrelation, and hence significantly outperforms either original algorithm

on difficult multimodal sampling problems.

2

Bayesian Hierarchical Modeling for Change Points in a Point Process

2.1 Introduction

Point processes are a class of random processes whose realizations consist of sets of isolated points in some space, such as the real line (often representing time) or the plane (representing geographical locations). Point processes are well studied and powerful tools for modeling spatial data (Daley, 1988; Diggle, 2003). It has also been applied to a variety of disciplines such as forestry (Penttinen and Stoyan, 2000), plant ecology (Law et al., 2009), epidemiology (Gatrell et al., 1996), seismology (Ogata, 1999), astronomy (Scargle and Babu, 2003), telecommunications (Eden et al., 2004), economics (Engle and Lunde, 2003), and others.

In this chapter, we use point process to analyze a particular volcanic dataset collected from Soufriére Hills volcano. The Soufriére Hills volcano is a complex stratovolcano located on the island of Montserrat, a British overseas territory. After a long period of dormancy, the volcano became active in 1995, and has continued to erupt ever since. It is well known that the most common and devastating result of an

explosive volcanic eruption is a pyroclastic flow (PF), which is a fast-moving current of hot gas and rock that can reach velocity as great as 450 mph moving away from the volcano (USGS, http://pubs.usgs.gov/gip/msh/pyroclastic.html). Pyroclastic flows normally travel downhill or spread laterally under gravity. Their speed depends upon the density of the current, the volcanic output rate, and the gradient of the slope. Pyroclastic flows can be extremely disastrous and fatal due to their high temperature and mobility. Because of the existing volcanic dome and the associated potential for pyroclastic activity, it is too dangerous to live in many parts of the island and the financial burden of relocating all the residents is too high to take. To this end, a group of geologists, statisticians, mathematicians, and physicists initiated a research project to construct the hazard map of the island (see Figure 2.1), i.e., to determine the probability of the occurrence of certain catastrophic events at any spatial location and any time instance. The goal of this chapter is to model the frequency of large pyroclastic flows with runout greater than 500 meters, under our assumption that the system is stationary.

The organization of the chapter is as follows. Section 2.2 describes the composition of the dataset. A visualization of the dataset indicates the existence of multiple change points in this process. Section 2.3 applies two existing approaches to analyze the dataset, namely, Bayesian Binary Segmentation (BBS) (Young and Kuo, 2001) and Reversible Jump Markov Chain Monte Carlo (RJMCMC) (Green, 1995). Section 2.4 proposes the Bayesian Hierarchical (BH) model to address the limitations of the BBS and RJMCMC approaches on this dataset. Section 2.5 describes procedures to make predictions from the Bayesian hierarchical model.

2.2 Pyroclastic flow data

Following a three-year period of heightened volcano-seismic activity beneath the island, the onset of phreatic volcanic activity started in July 1995 at the Soufriére Hills

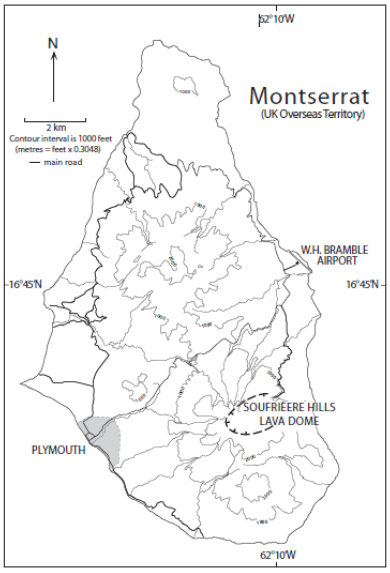


FIGURE 2.1: The island of Montserrat. Adapted with permission from Calder et al. 2002.

volcano, Montserrat on 18 July 1995. Phreatic explosions gave way to continuous eruption of juvenile andesitic magma in the form of a lava dome on or around 15 November 1995 (Young et al., 1998). The dataset to be analyzed in this chapter contains a list of dates between March 27, 1996 and July 28, 2008 when pyroclastic flows greater than 500 meters runout occurred. There were 868 such events in a time period of 4507 days. The cumulative count of pyroclastic flows is depicted in Figure 2.2, in which each point represents an event. If there are multiple events in a short time period, the dots are connected as a line with a steep slope. Otherwise, the dots are scattered. It is easy to identify different rates of increase during the course of observation. Hence, we would like to make inference and prediction on the change points of the dataset. There are several existing methods in the literature to deal with (multiple) change points (Andrews, 1993; Bai, 1997; Raftery and Akman, 1986; Chib, 1998; Brodsky and Darkhovsky, 1993; Green, 1995). In the next section, we first consider Bayesian Binary Segmentation approach for some preliminary analysis of the dataset.

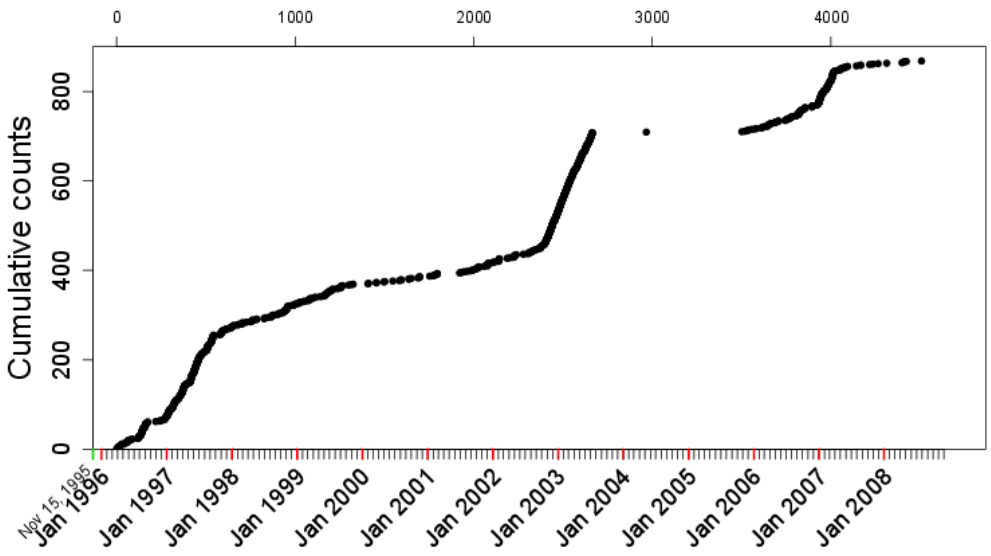


FIGURE 2.2: The cumulative counts of pyroclastic flows with greater than 500 meters runout between years 1996 and 2008.

2.3 Existing models for multiple change points

2.3.1 Bayesian binary segmentation

In this subsection, we apply the Bayesian binary segmentation (BBS) method (Young and Kuo, 2001) to perform some preliminary analysis of the dataset. The BBS method describes a series of model selection procedures, with the advantage of easy implementation. Specifically, in each step, one assumes two models: A no-changepoint model H_0 versus a single-changepoint model H_1. If the Bayes factor

$$\frac{p(H_1 \mid \text{data})}{p(H_0 \mid \text{data})} > 1,$$

model H_1 is selected and then change points and rates are estimated accordingly.

We apply the BBS method to the given dataset and show the estimated change points and rates in Figure 2.3. Though easy to implement, the results reveal two major drawbacks of the BBS method. First, prior information is only used to assign equal prior probability to the two models— it offers no way to include prior

information about the frequency of change points or the magnitude of the changes. Second, the method uses a very coarse grain threshold in detecting change point and the algorithm can be very greedy. Notice that in each step, model H_1 is selected and one more change point is detected if the Bayes factor exceeds one. However, with our dataset, we find that in many steps, the Bayes factor is very close to one. Without procedures to discriminate the subtle difference, the BBS method detects an unreasonably large number of changepoints.

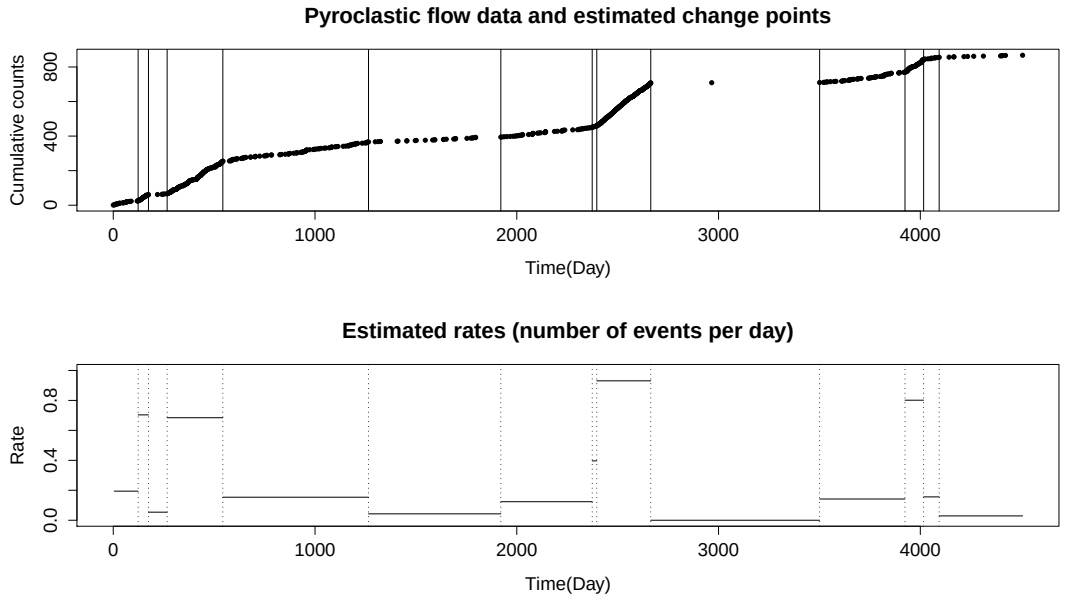


FIGURE 2.3: Estimated change points and rates of the pyroclastic flow dataset obtained by Bayesian binary segmentation method.

2.3.2 Reversible jump Markov chain Monte Carlo

The Reversible Jump Markov Chain Monte Carlo (RJMCMC) algorithm was developed by Peter Green in his attempt to analyze and detect multiple change points in the rate of coal mining disasters (Green, 1995).

When events occur continuously and independently, it is natural to model them as an inhomogeneous Poisson process. Green chose to model the coal mining events

as a Poisson process with a step rate function $\lambda(t)$ as illustrated in Figure 2.4. In the plot, L is the length (in days) of the entire time period, $s_1, s_2, \ldots, s_k$ are the times of change points, and k is the number of change points. The function $\lambda(t)$ remains constant in the period between two successive change points, and then steps up or down. Therefore, the parameters in the model include: 1) the number of change points k; 2) the position of change points $\vec{s} = \{s_1, \ldots, s_k\}$, and 3) the expected number of events that occur per day $\{\lambda_j\}_{j=1}^{k+1}$. In the following, we denote the whole set of parameters by Θ:

$$\Theta = (k,\ s_1, \ldots, s_k,\ \lambda_1, \ldots, \lambda_{k+1}).$$

Notice that the number of change points k is itself a parameter. As a result, the dimension of the parameter space is unknown and regular Markov chain Monte Carlo (MCMC) is not applicable to this case.

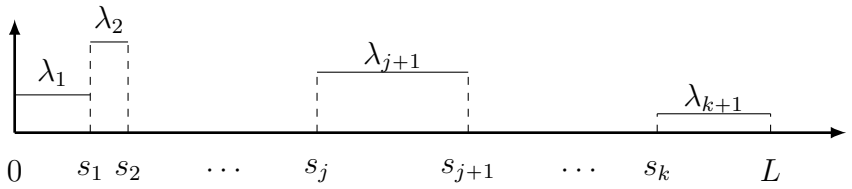


FIGURE 2.4: An illustration of a Poisson process with a step rate function.

The most common choices of prior distributions for parameters representing counts (like k) and positive-valued parameters (like $\{\lambda_j\}$ are the Poisson and Gamma distributions, respectively. To make the change points more spread out, Green modeled them as the even-numbered order statistics of $2k+1$ uniformly distributed points for $\vec{s}$. Specifically, the prior distributions are summarized as follows:

1. $k \sim \text{Poi}(L\delta_0)$, where δ_0 is the fixed frequency of rate change per day. Then $L\delta_0$ is the prior expected number of change points in the whole time period $[0, L]$.

2. Positions of the change points are the even numbered order statistics of $2k + 1$

points uniformly distributed on $[0, L]$, to discourages very short periods between successive change points. Thus:

$$\pi(s_1, s_2, \ldots, s_k \mid k) = \frac{(2k+1)!}{L^{2k+1}} \prod_{i=1}^{k+1} (s_i - s_{i-1}) \mathbf{1}_{\{0<s_1<\cdots<s_k<L\}},$$

where $s_0 = 0$ and $s_{k+1} = L$.

3. $\{\lambda_1, \lambda_2, \ldots, \lambda_{k+1} \mid k\} \overset{\text{idd}}{\sim} \text{Gamma}(\alpha, \beta)$.

The likelihood function is given by

$$\ell(\tau_1, \tau_2, \ldots, \tau_N \mid k, s_1, \ldots, s_k, \lambda_1, \ldots, \lambda_{k+1}) = \prod_{i=1}^{k+1} \lambda_i^{n_i} e^{-\lambda_i (s_i - s_{i-1})},$$

where $\tau_1, \tau_2, \ldots, \tau_N$ denote the dates of N events, and n_i is the number of events in the time interval $(s_{i-1}, s_i]$.

In each RJMCMC step, one of four types of moves might be implemented. They are *rate change*, *position change*, *birth* and *death*, each proposed with a specified probability. Suppose that there are k change points in the previous iteration ($2k + 2$ parameters); then "rate change" is to choose any of the $k + 1$ rates and propose a change by applying Gibbs sampling method (Gelfand and Smith, 1990; Casella and George, 1992). Similarly, "position change" is to modify the position of one change points and is achieved by Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970). A "birth" move is to add a new change point and subsequently two more new rates, while a "death" move is the reverse of the birth move— the removal of a change point and merger of the periods that had preceded and followed it. Both birth and death can be implemented as Metropolis-Hastings steps. Figures 2.5 and 2.6 describe the change of the parameter sets before and after the two moves, respectively.

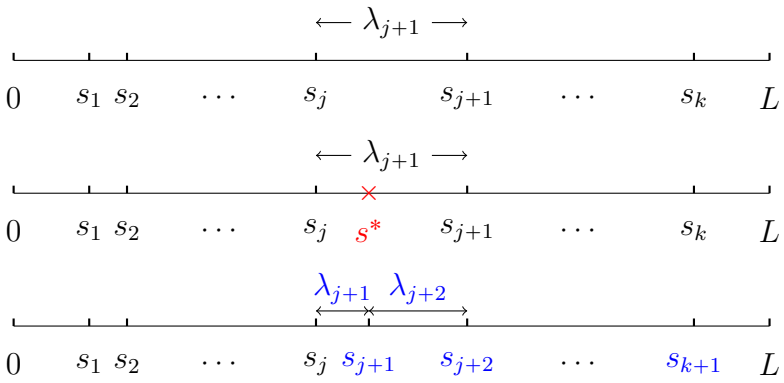


FIGURE 2.5: Illustration of Birth Move. A new change point s^* is proposed between the original change points s_j and s_{j+1}. Two new rates are proposed before and after s^*.

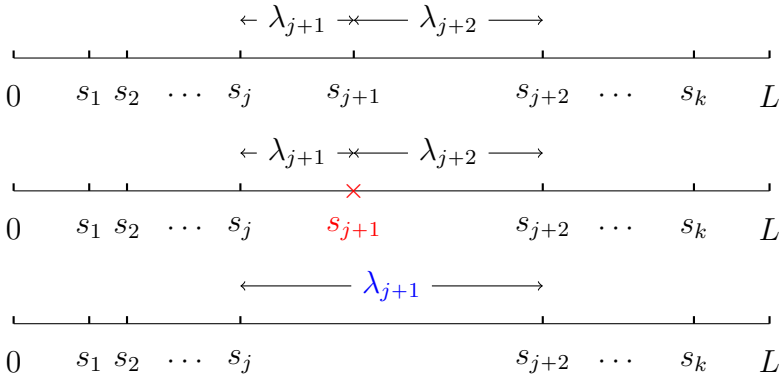


FIGURE 2.6: Illustration of Death Move. A change point s_{j+1} is removed and there is a single rate between s_j and s_{j+2}.

We apply the reversible jump MCMC method to the pyroclastic flow data. The probability distribution for the rate change, position change, birth move and death move are chosen to be $pR = 1/6$, $pP = 1/6$, $pB = 1/3$, and $pD = 1/3$. Birth and death moves are attempted more frequently because the acceptance rates for these two moves are usually low. Green chose $\alpha = 1$ and $\beta = 200$ for the Gamma distribution as the prior of rates to achieve a prior mean α/β close to the empirical mean of 192disasters/40907 days. Our data set features $N = 868$ events in $L = 4507$ days, and accordingly we set $\alpha = 1$ and $\beta = 5$ in our study. We run chains of length 100,000 steps, and show posterior results summarized from 8,000 samples (first

thinning for every 10 original MCMC samples, and then use a burn-in period 2,000). Table 2.1 shows the relationship between the posterior mode of k—the number of change points—and prior means. As $L\delta$ increases, so does the posterior mode of k. The sensitivity of the posterior result is easy to identify. Additionally, the fixed prior parameters are not suitable for the next step of our analysis, prediction.

Table 2.1: Posterior modes of the number of change points for different priors.

prior mean ($L\delta$)	5	10	15	20
posterior mode (k)	12	14	15	18

2.4 Bayesian hierarchical model

In this section, we discuss several strategies to improve upon the existing BBS and RJMCMC methods.

2.4.1 Hierarchical model with objective priors

The first improvement is to introduce a second level of hierarchy with hyper-parameters. This means, δ, α and β are not fixed parameters. Instead, they are assigned objective hyper-prior distributions:

$$\begin{aligned} \delta &\sim \pi(\delta) \propto 1/\sqrt{\delta} 1_{[0,1]}, \\ (k \mid \delta) &\sim \text{Poi}(L\delta), \\ \alpha &\sim \pi(\alpha) \propto 1 \text{ on } (0, \infty) \\ \beta &\sim \pi(\beta) \propto 1 \text{ on } (0, \infty). \end{aligned}$$

We constrain δ to the range $[0, 1]$, at most one change-point per day on average. It follows that the prior mean for k will be in the range of $[0, L]$. Despite the improper priors for α and β, the posterior will be proper unless $k = 0$, i.e., unless there is no change point. Because we believe the frequency of large pyroclastic flows is changing

with the dynamic activities of the volcano, we find the model with $k = 0$ to be untenable and set its likelihood to zero, ensuring posterior propriety.

Next, since Green's even-order-statistics prior for $\vec{s}$ made little difference in the results for pyroclastic flow data, for the convenience of prediction, we instead modeled $\vec{s}$ as the ordered statistics of uniformly variables on $[0, L]$, to make $\{s_j\}$ the events in a homogeneous Poisson process:

$$(s_1, \ldots, s_k \mid k) \sim \text{Unif}(0, L).$$

Prior distributions for the rates remain the same. To accommodate the introduction of three more parameters, two more moves are also required. By conjugacy, Gibbs sampling suffices to update δ and β, while a Metropolis-Hastings step is applied to update α.

We first apply this model to a simulated time series of pyroclastic flows. The simulation included $k = 9$ change points, with $N = 2111$ events in $L = 4500$ days. We apply the proposed hierarchical model to the simulated data and show posterior results summarized from 1,000 MCMC samples (100,000 original samples, thinning 50 and then burn-in period 1,000). Figure 2.7a shows the posterior distribution of k, the number of change points, and the it shows in the histogram that the posterior mode of k is 14, although the true value is 9. Figure 2.7b depicts both the true positions of the change points in the simulated pyroclastic flow data and the posterior samples of change point positions. The plot on the top of the figure shows the cumulative counts of simulated pyroclastic flows (in black) and where changes of frequency occur in our design (red vertical dashed lines). The plot in the bottom of the figure shows a collection of posterior samples of change point positions. In each line horizontally, the black dots indicate the change point positions for one of those iterations with posterior number of change points equal to 14 (posterior mode). When dots form a vertical line at certain positions, it implies that the sampler recognizes those positions

as change points with a large degree of certainty. In contrast, those scattered points imply some positions might be considered as change points but yet the sampler is not very certain about that. Those vertical lines happen to match the positions of the real change points. Figure 2.7c displays posterior mean rate and its 90% credible intervals together with the real rates. We can see that the posterior means are extremely close to the true values.

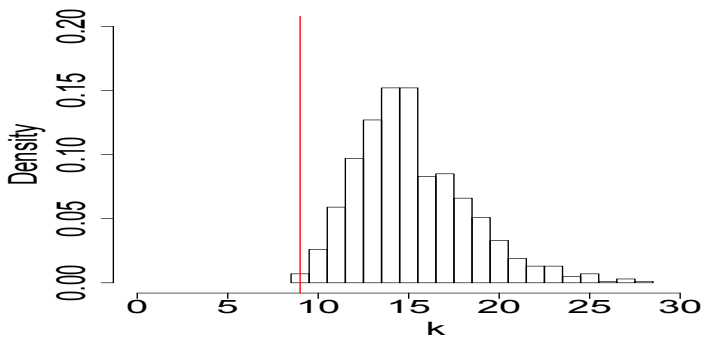
The overall results of change point detection and rate estimation are good. However, we can observe from the figures that the hyper-prior distributions still resulted in an excessive number of change points. And from the posterior samples, we also plot the posterior rate from one single MCMC sample with a relatively large posterior number of change points in Figure 2.8. It turns out that some adjacent rates are quite close to each other in value. This is not favorable because we would like to detect obvious changes in the eruption frequency, although neither the prior nor the likelihood embodies this preference by penalizing tiny changes.

2.4.2 Penalized prior distribution

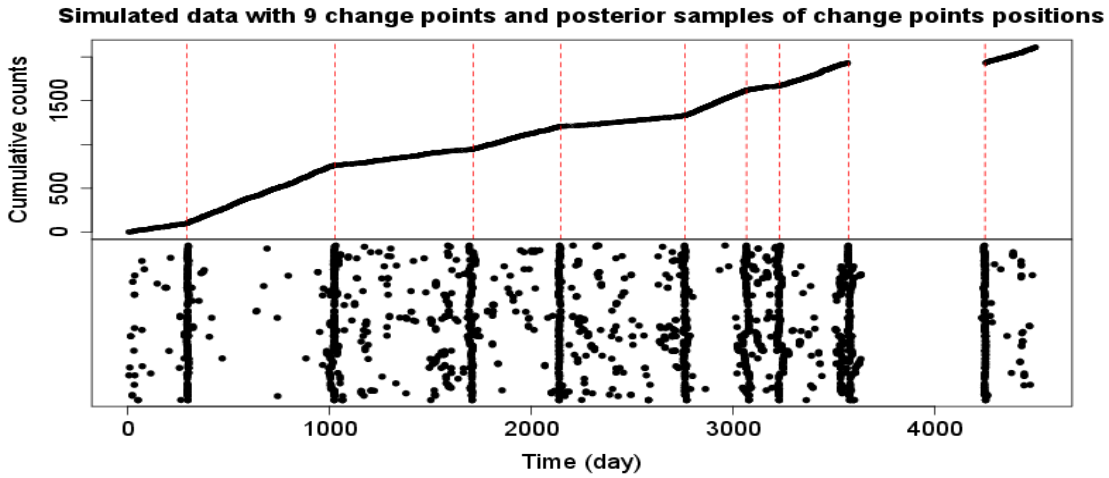
In our second improvement, we modify the prior for rates by adding a penalty term to the Gamma density to constrain on close consecutive rates:

$$\pi(\lambda_1, \ldots, \lambda_{k+1} \mid k, \alpha, \beta) = \frac{1}{C_k^{(\alpha,\beta,\phi)}} \prod_{i=1}^{k+1} (\lambda_i)^{\alpha-1} e^{-\beta\lambda_i} \prod_{i=1}^{k} \mid \lambda_{i+1} - \lambda_i \mid^{\phi},$$

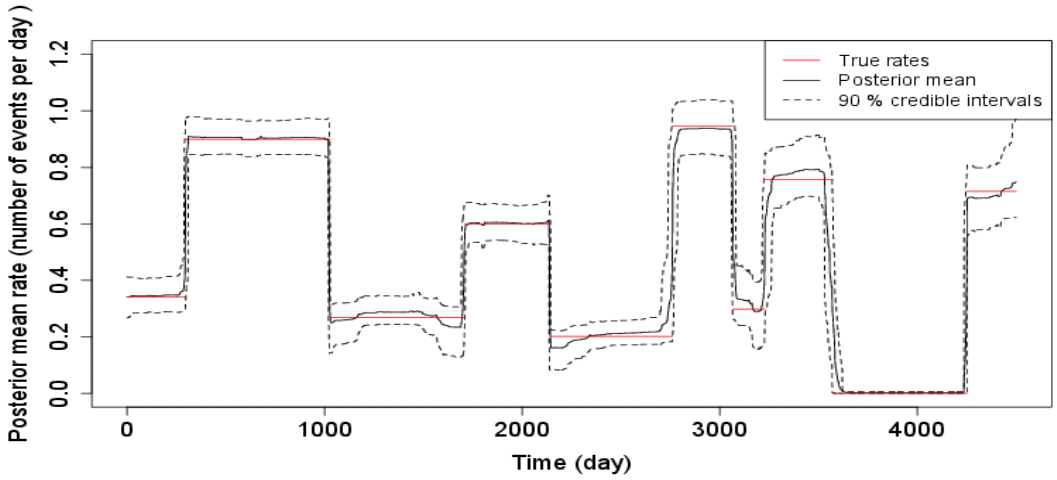
where the nonnegative parameter ϕ controls the degree of penalty. If any one of the adjacent pair of rates are close, the density decreases. The larger ϕ is, the greater the penalty. This new prior leads to one extra difficulty: the normalizing constant $C_k^{(\alpha,\beta,\phi)}$, a $(k+1)$-dimensional integral that depends on parameters k, α, β and ϕ, cannot be expressed in closed form except for a few special cases. For instance, when



(a) Posterior distribution of the number of change points.



(b) True positions of the change points in the simulated pyroclastic flow and the posterior change point positions.



(c) Posterior mean rate and its 90% credible intervals with the real rates.

FIGURE 2.7: Results of simulation study I.

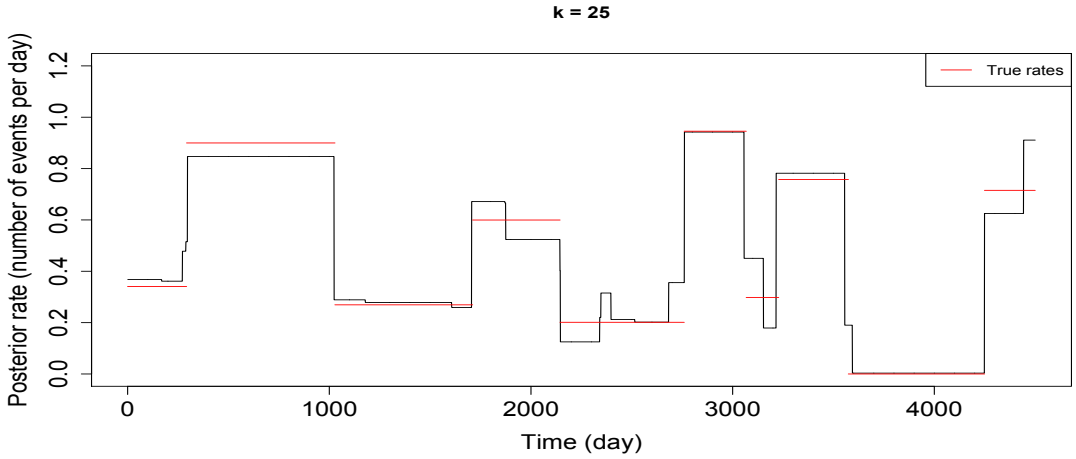


FIGURE 2.8: A single MCMC sample with a relatively large posterior number of change points.

$\phi = 2$, the normalizing constant has the following form:

$$C_k^{(\alpha,\beta,\phi=2)} = \int (\lambda_1 \cdots \lambda_{k+1})^{\alpha-1} e^{-\beta(\lambda_1+\cdots+\lambda_{k+1})} (\lambda_2 - \lambda_1)^2 \cdots (\lambda_{k+1} - \lambda_k)^2 \, d\lambda_1 \cdots d\lambda_{k+1},$$

which can be solved recursively:

$$\begin{aligned}
4C_{k+1} &= \frac{\Gamma(\alpha+2)}{\beta^{\alpha+2}} C_k - 2\frac{\Gamma(\alpha+1)}{\beta^{\alpha+1}} A_k + \frac{\Gamma(\alpha)}{\beta^{\alpha}} B_k, \\
A_k &= \frac{\Gamma(\alpha+3)}{\beta^{\alpha+3}} C_{k-1} - 2\frac{\Gamma(\alpha+2)}{\beta^{\alpha+2}} A_{k-1} + \frac{\Gamma(\alpha+1)}{\beta^{\alpha+1}} B_{k-1}, \\
B_k &= \frac{\Gamma(\alpha+4)}{\beta^{\alpha+4}} C_{k-1} - 2\frac{\Gamma(\alpha+3)}{\beta^{\alpha+3}} A_{k-1} + \frac{\Gamma(\alpha+2)}{\beta^{\alpha+2}} B_{k-1},
\end{aligned}$$

with initial values:

$$\begin{aligned}
A_1 &= \frac{\Gamma(\alpha)\Gamma(\alpha+3) - \Gamma(\alpha+1)\Gamma(\alpha+2)}{\beta^{2\alpha+3}}, \\
B_1 &= \frac{\Gamma^2(\alpha+2) - 2\Gamma(\alpha+1)\Gamma(\alpha+3) + \Gamma(\alpha)\Gamma(\alpha+4)}{\beta^{2\alpha+4}}, \\
C_0 &= \frac{\Gamma(\alpha)}{\beta^{\alpha}} \quad \text{and} \quad C_1 = \frac{2[\Gamma(\alpha)\Gamma(\alpha+2) - \Gamma(\alpha+1)]^2}{\beta^{2\alpha+2}}.
\end{aligned}$$

However, for $\phi < 2$, we can only compute the normalizing constant approximately, for example, by applying Monte Carlo integration. In this case, the normalizing constant takes the form:

$$C_k^{(\alpha,\beta,\phi)} = \int (\lambda_1 \cdots \lambda_{k+1})^{\alpha-1} \, e^{-\beta(\lambda_1+\cdots+\lambda_{k+1})} \mid \lambda_2 - \lambda_1 \mid^{\phi} \cdots \mid \lambda_{k+1} - \lambda_k \mid^{\phi} \, d\,\lambda_1 \cdots d\,\lambda_{k+1}.$$

By changing variables, we may take $\beta = 1$ (and compute $C_k^{(\alpha,1,\phi)}$) without loss of generality, since we can then retrieve:

$$C_k^{(\alpha,\beta,\phi)} = \frac{C_k^{(\alpha,1,\phi)}}{\beta^{(\alpha+\phi)(k+1)-\phi}},$$

and evaluate the ratio of successive normalizing constants

$$\frac{C_k^{(\alpha,\beta,\phi)}}{C_{k+1}^{(\alpha,\beta,\phi)}} = \frac{C_k^{(\alpha,1,\phi)}}{C_{k+1}^{(\alpha,1,\phi)}} \beta^{\alpha+\phi}.$$

The latter will be needed in the computation of acceptance probability, particularly prior ratios.

A routine Monte Carlo importance sampling scheme is to draw $\lambda_i^{(m)} \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, 1)$, for $i = 1, \ldots, k+1$ and $m = 1, \ldots, M$, and estimate the normalizing constant by

$$\widehat{C}_k^{(\alpha,1,\phi)} = \frac{1}{M} \sum_{m=1}^{M} \Gamma(\alpha)^{k+1} \mid \lambda_2^{(m)} - \lambda_1^{(m)} \mid^{\phi} \cdots \mid \lambda_{k+1}^{(m)} - \lambda_k^{(m)} \mid^{\phi},$$

and the ratio by

$$\widehat{\frac{C_{k-1}^{(\alpha,1,\phi)}}{C_k^{(\alpha,1,\phi)}}} = \frac{\widehat{C}_{k-1}^{(\alpha,1,\phi)}}{\widehat{C}_k^{(\alpha,1,\phi)}},$$

By the law of large numbers, the sample average of N replications of these estimates should converge to the quantities of interest. However, the performance of the approximation is not satisfactory. While the numerical results of the recursive

algorithm show convergence of the normalizing constants ratio for $\phi = 2$, there is obvious oscillation in the results obtained by the regular Monte Carlo integration method, which gets more visible for larger ϕ. The reason for the poor performance of the Monte Carlo integration is that the normalizing constant is a $k+1$ dimensional integral, and as k increases, the variance for the Monte Carlo estimate becomes larger and larger, and thus the estimates for the normalzing constant (and ratios) are not reliable.

We now propose a new Monte Carlo integration method using a dimension reduction strategy. Specifically, denote

$$f(\lambda_1, \ldots, \lambda_{k+1}) = (\lambda_1 \cdots \lambda_{k+1})^{\alpha-1} e^{-(\lambda_1+\cdots+\lambda_{k+1})} \mid \lambda_2 - \lambda_1 \mid^{\phi} \cdots \mid \lambda_{k+1} - \lambda_k \mid^{\phi},$$

and

$$d\,\lambda^{(k+1)} = d\,\lambda_1 \cdots d\,\lambda_{k+1}.$$

Then the normalizing constants ratio is:

$$\begin{aligned}\frac{C_{k+1}^{(\alpha,1,\phi)}}{C_k^{(\alpha,1,\phi)}} &= \frac{\int f(\lambda_1, \ldots, \lambda_{k+1}) \lambda_{k+2}^{\alpha-1} e^{-\lambda_{k+2}} \mid \lambda_{k+2} - \lambda_{k+1} \mid^{\phi} d\,\lambda^{(k+1)} d\,\lambda_{k+2}}{\int f(\lambda_1, \ldots, \lambda_{k+1}) d\,\lambda^{(k+1)}} \\ &= \int f^*(\lambda_1, \ldots, \lambda_{k+1}) \lambda_{k+2}^{\alpha-1} e^{-\lambda_{k+2}} \mid \lambda_{k+2} - \lambda_{k+1} \mid^{\phi} d\,\lambda^{(k+1)} d\,\lambda_{k+2}.\end{aligned}$$

Now the ratio of two $k+1$ dimensional integrals becomes a one-dimensional integral, which greatly reduces the potential variance. One estimate of the normalizing constants ratio is obtained by

$$\widehat{\frac{C_{k+1}^{(\alpha,1,\phi)}}{C_k^{(\alpha,1,\phi)}}} = \frac{1}{M} \sum_{m=1}^{M} \Gamma(\alpha) \mid \lambda_{k+2}^{(m)} - \lambda_{k+1}^{(m)} \mid^{\phi},$$

where $\{\lambda_{k+2}^{(m)}\}_{m=1}^{M}$ are independently drawn from Gamma$(\alpha, 1)$ and $\{\lambda_{k+1}^{(m)}\}_{m=1}^{M}$ are samples from the marginal distribution $f^*(\lambda_{k+1})$.

We can draw samples $\{\lambda_{k+1}^{(m)}\}_{m=1}^{M}$ from the joint distribution $f(\lambda_1, \ldots, \lambda_{k+1})$ using one-step-at-a-time-Gibbs-Metropolis. At the mth iteration, we update each λ_j twice in a cycle. The procedure has two steps. First, update $\lambda_j^{(m-1)}$ for each $j = 1, 2, \ldots, k+1$, by proposing a new sample λ_j^* from Gamma(α,1) and accepting it with probability

$$\min\left\{1, \frac{f(\lambda_1^{(m)}, \ldots, \lambda_{j-1}^{(m)}, \lambda_j^*, \lambda_{j+1}^{(m-1)}, \ldots, \lambda_{k+1}^{(m-1)})}{f(\lambda_1^{(m)}, \ldots, \lambda_{j-1}^{(m)}, \lambda_j^{(m-1)}, \lambda_{j+1}^{(m-1)}, \ldots, \lambda_{k+1}^{(m-1)})}\right\}.$$

Second, for each $j = k, \ldots, 1$, propose λ_j^* from Gamma(α,1) and accept it with probability

$$\min\left\{1, \frac{f(\lambda_1^{(m)}, \ldots, \lambda_{j-1}^{(m)}, \lambda_j^*, \lambda_{j+1}^{(m)}, \ldots, \lambda_{k+1}^{(m)})}{f(\lambda_1^{(m)}, \ldots, \lambda_{j-1}^{(m)}, \lambda_j^{(m)}, \lambda_{j+1}^{(m)}, \ldots, \lambda_{k+1}^{(m)})}\right\}.$$

Figure 2.9 plots $\frac{\widehat{C_{k+1}^{(\alpha,1,\phi)}}}{C_k^{(\alpha,1,\phi)}}$ for $\alpha = 1$ and $\beta = 1$ particularly, at a set of k values for each of three ϕ values. When $\phi = 2$, we can compare the numerical approximation with the exact results. All three curves show convergence when k increases.

In the remaining of this section, we study the simulated pyroclastic flow data using the new prior distribution for rates and choose prior distributions for α according to ϕ values. The prior distributions for other parameters are unchanged. For $\phi = 0.5$, the normalizing constants ratios are estimated by Monte Carlo integration. They are pre-calculated before MCMC runs to avoid time-consuming inline computation. We assign a discrete uniform prior distribution for α:

$$\pi(\alpha) = \text{Unif}\,\{0.1, 0.2, 0.3, \ldots, 2.9, 3.0\},$$

for computational efficiency.

In Figure 2.10a, the histogram shows that the posterior mode of k is 12, still larger than the true value 9 but smaller than that in the previous simulation study. This

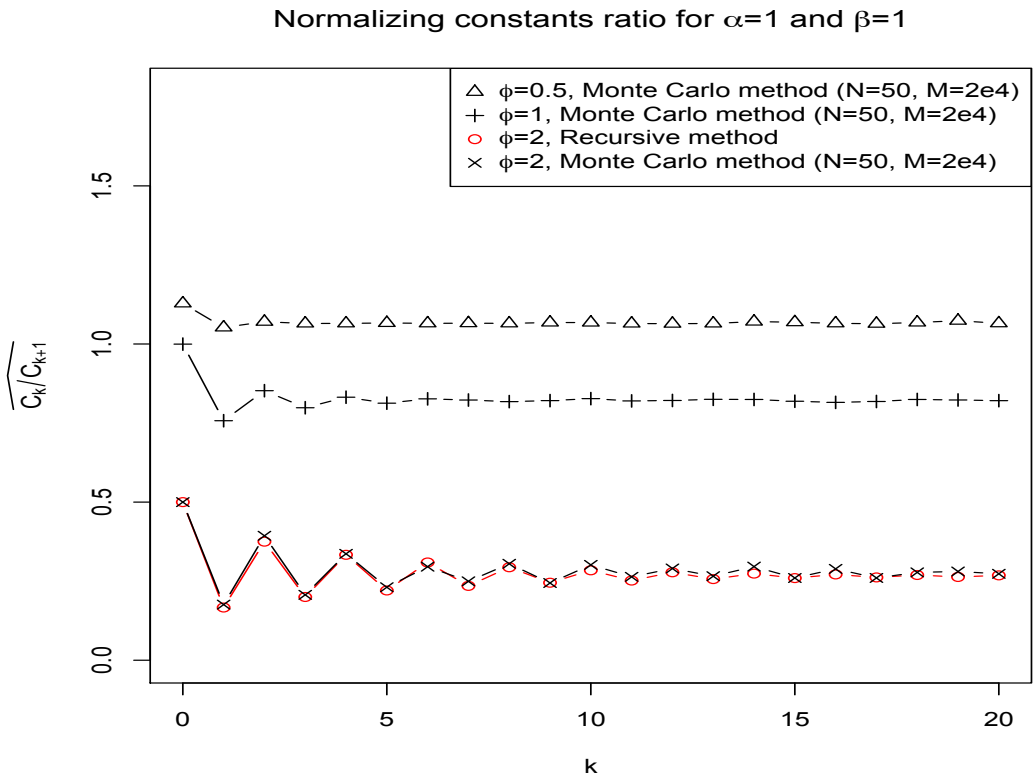
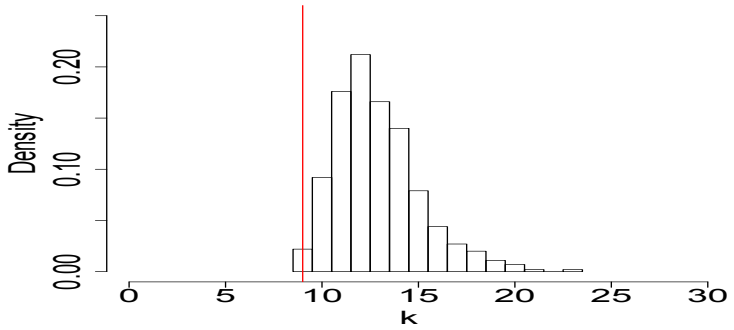


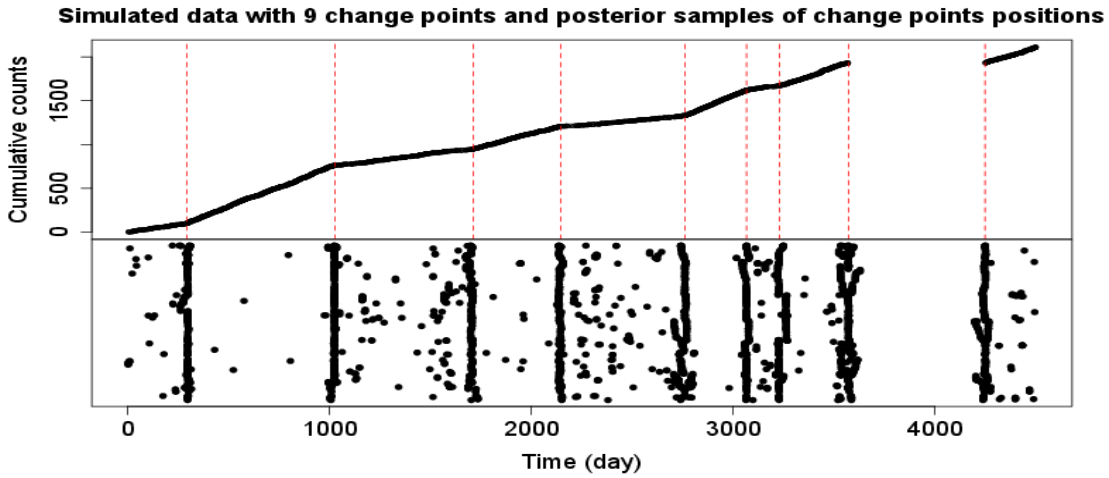
FIGURE 2.9: Normalizing constants ratio for specific α and β values, and three ϕ values, computed with Monte Carlo method and recursive method.

is due to the penalty term as it adds the constraint to close values of adjacent rates. Figures 2.10b and 2.10c demonstrate posterior results for the change point positions and rates, which match the true values pretty well and has smaller discrepancies than the one in Section 2.4.1.

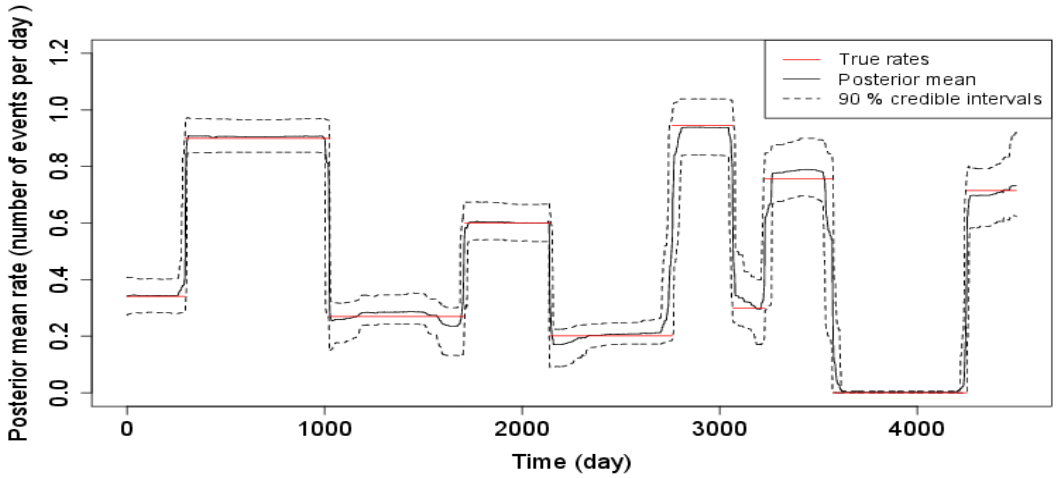
The normalizing constants ratio can be calculated exactly and quickly for the case $\phi = 2$, by recursion. As a result, we can continue using the objective prior for α: $\pi(\alpha) \propto 1$ on $(0, \infty)$. The posterior distribution shown in Figure 2.11a indicates that the posterior mode of the number of change points is 9, exactly the true value. In addition, the MCMC draws of the change point positions plotted in Figure 2.11b find the all the true positions correctly, yet with certain degree of uncertainty at some places shown by wiggles along the lines. In addition, the posterior mean rate



(a) Posterior distribution of the number of change points.



(b) Poster distribution of $\boldsymbol{s}$.



(c) Posterior mean rate.

FIGURE 2.10: Results of simulation study II, under the penalty prior for rates with $\phi = 0.5$.

shown in Figure 2.11c is still a good estimate compared to the true rates.

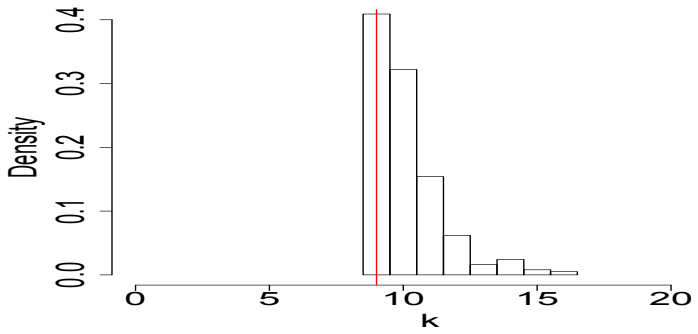
The simulation study demonstrates good performance of the reversible jump MCMC sampler with a penalized prior on rates and positive penalty ϕ, in terms of correctly identifying the true number and positions of the change points and closely estimating the rates as well. Therefore, we end this section by applying the above method to the real pyroclastic flow dataset. Figures 2.12a–2.12c show the posterior results. The histogram for k indicates a range from 9 to 14 with the mode at 11, and the posterior samples of change point positions mostly form into 11 lines. The relatively straight lines imply absolute changes while those wiggling lines imply some uncertainty.

2.4.3 Mixture prior distribution

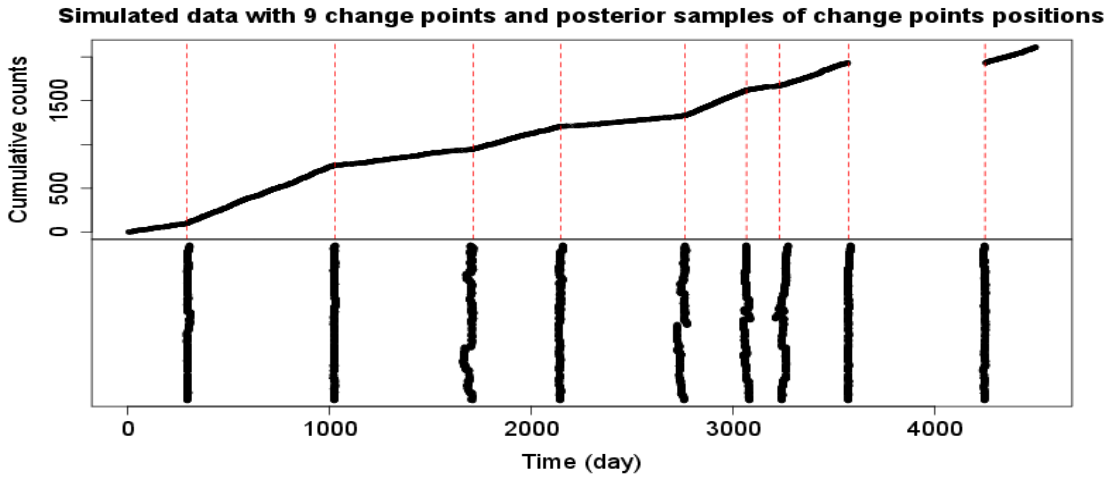
Our volcanologist collaborators report that the rates can be exactly zero when the lava dome stops erupting and enters an inactive/dormant period. However, in our previous analysis, we have assumed that rates are always positive. To improve our model in this aspect, we would like to add a point mass at zero to the rates in the prior distribution. We first introduce latent variables $\{z_i\}_{i=1}^{k+1}$ indicating whether rate λ_i is positive and define

$$\lambda_i = z_i \lambda_i^* + (1 - z_i) 1_{\{0\}}, \quad i = 1, \ldots, k+1.$$

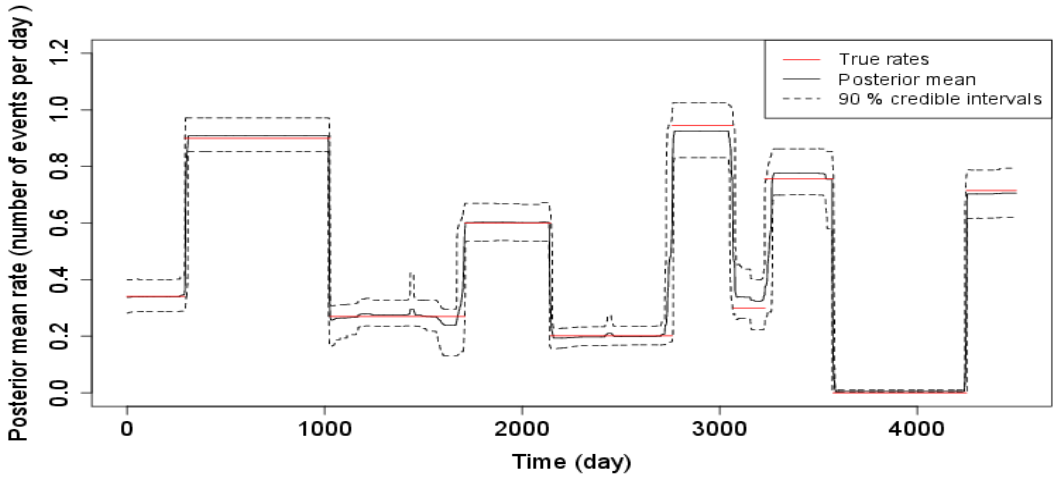
That is, λ_i is positive when $z_i = 1$ and zero when $z_i = 0$. We assign independent Bernoulli distribution with success probability p to each z_i, and a hyper-prior distribution Beta(1,1) (or equivalently Unif(0,1)) to p. The joint prior distribution of



(a) Posterior distribution of the number of change points.

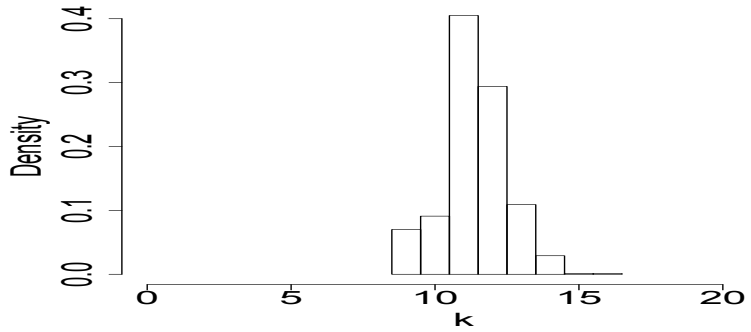


(b) Poster distribution of $\boldsymbol{s}$.

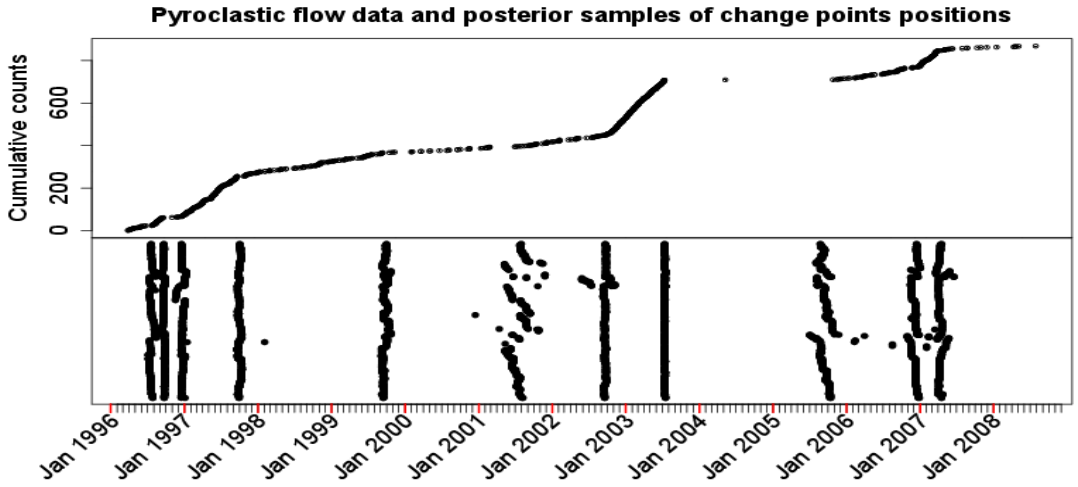


(c) Posterior mean rate.

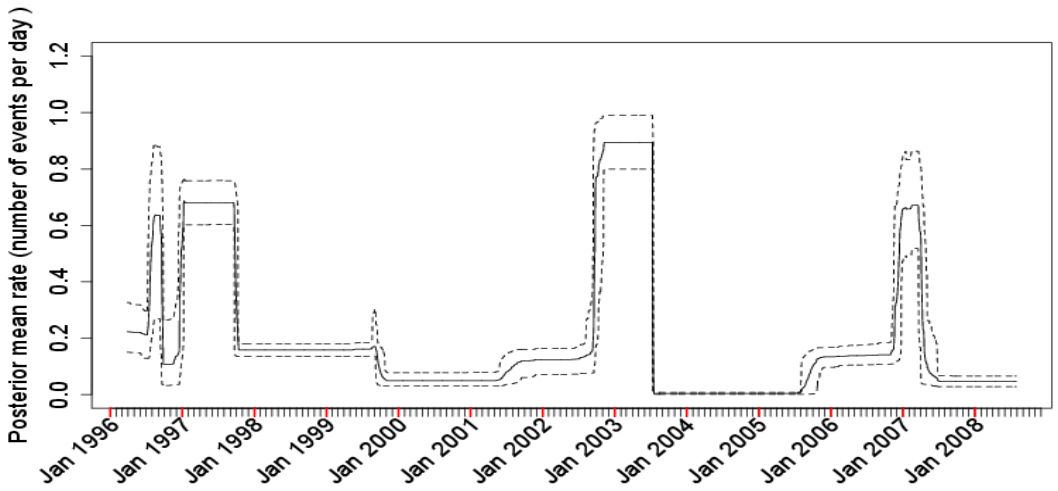
FIGURE 2.11: Results of simulation study II, under the penalty prior for rates with $\phi = 2$.



(a) Posterior distribution of the number of change points.



(b) Poster distribution of $\boldsymbol{s}$.



(c) Posterior mean rate.

FIGURE 2.12: Real data analysis, under the penalty prior for rates with $\phi = 2$.

$p, \lambda_1^*, \ldots, \lambda_{k+1}^*, z_1, \ldots, z_{k+1}$ conditional on k, α, β, and ϕ is:

$$\begin{aligned}
&\pi(p, z_1, \ldots, z_{k+1}, \lambda_1^*, \ldots, \lambda_{k+1}^* \mid k, \alpha, \beta, \phi) \\
=&\frac{1}{C_k^{(\alpha,\beta,\phi)}} \pi(p)\pi(z_1, \ldots, z_{k+1} \mid p, k) \prod_{i=1}^{k+1} (\lambda_i^*)^{\alpha-1} e^{-\beta\lambda_i^*} \prod_{i=1}^{k} \mid z_{i+1}\lambda_{i+1}^* - z_i\lambda_i^* \mid^{\phi} \\
=&\frac{1}{C_k^{(\alpha,\beta,\phi)}} \left[\prod_{i=1}^{k+1} p^{z_i}(1-p)^{1-z_i}\right] \cdot \left[\prod_{i=1}^{k+1} (\lambda_i^*)^{\alpha-1} e^{-\beta\lambda_i^*}\right] \cdot \left[\prod_{i=1}^{k} \mid z_{i+1}\lambda_{i+1}^* - z_i\lambda_i^* \mid^{\phi}\right],
\end{aligned}$$

where $C_k^{(\alpha,\beta,\phi)}$ is the normalizing constant. In this case, the normalizing constant or its ratio is too complicated to calculate analytically, and therefore is approximated by Monte Carlo integration methods. With routine Monte Carlo importance sampling, the normalizing constants ratio tends to fluctuate more dramatically for larger k ($k > 15$) or smaller α ($\alpha < 1$) values. However, with the new method presented in the previous section, the results are much more stable. Figure 2.13 show the estimates of the normalizing constant ratio for $\alpha = 0.5, 1$, $\beta = 1$ and $\phi = 0.5, 1, 2$ at different k values.

In the following, we apply the mixture prior distribution with $\phi = 2$ to the simulated dataset and present the posterior results in Figure 2.14. The posterior mode of k, posterior samples of $\{s_j\}$ and posterior mean of $\{\lambda_j\}$ are very good estimates of the corresponding parameters. A new plot shown in Figure 2.14b demonstrates the posterior probability of the rate being exactly zero at each time point. Since the sampler identifies the change points at correct moments, it also finds the large gap in the simulated dataset where the rate was simulated as zero. Similarly, we apply the new mixture prior distribution to the real data. Figures 2.15–2.18 plot all the posterior estimates of the parameters. We observe there is one event (even though it is not an outcome of the eruption) in the long gap and thus the posterior probability of the rate being zero during that time period is zero. To compare, we run the sampler to the real pyroclastic flow data without that point and plot the posterior

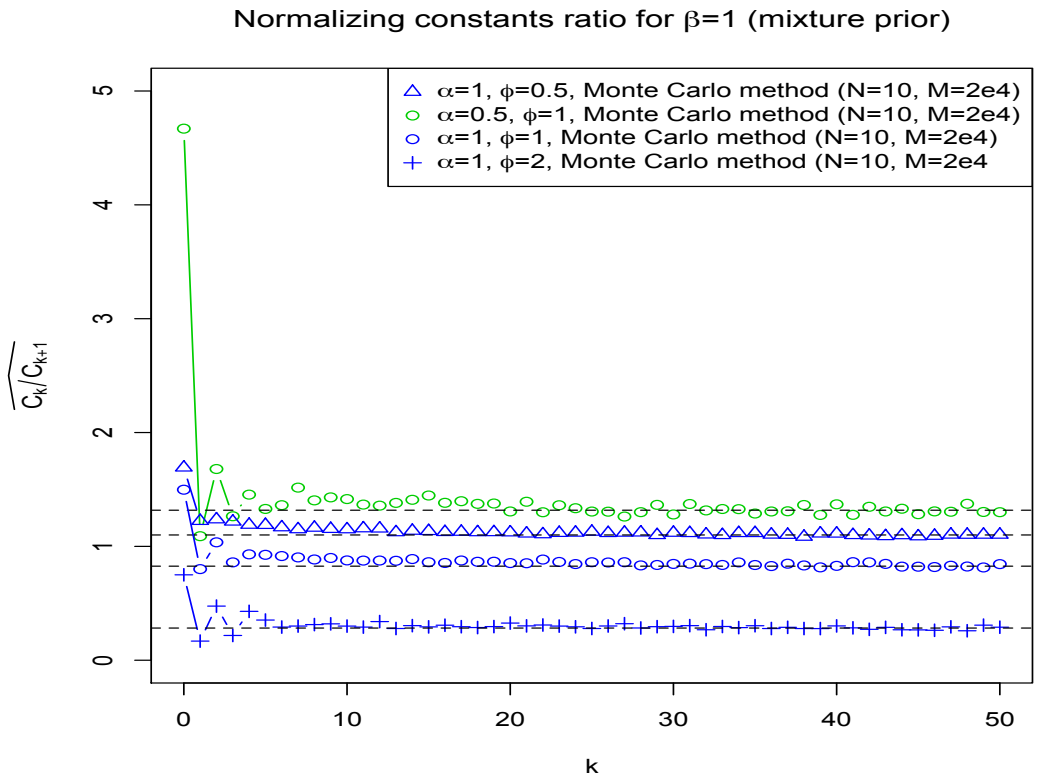


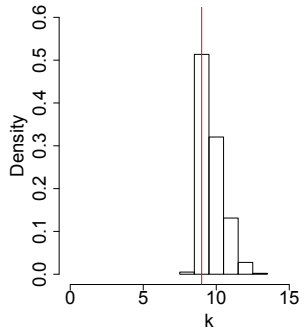
FIGURE 2.13: Mixture prior: normalizing constant ratio for different ϕ and α values.

probability again in Figure 2.19. Now, the sampler gives a much high probability on that interval that the rate equals zero.

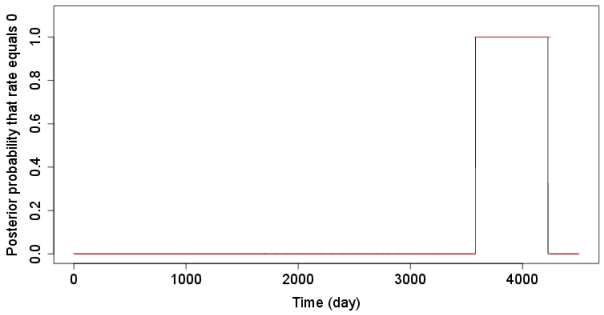
2.5 Prediction

Compared with Bayesian binary segmentation and the original reversible jump MCMC, the biggest advantage of the Bayesian hierarchical model is that it enables predictions based on what we have observed and the parameters we have estimated that accurately reflect all sources of uncertainty. In fact, in each RJMCMC replication, upon obtaining instantiations of all parameters:

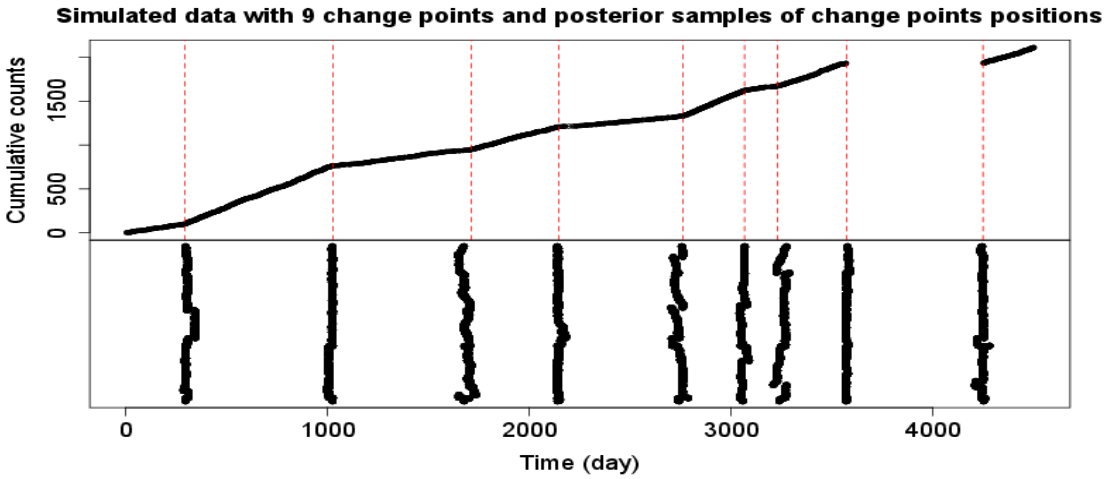
$$\widehat{\Theta} = (\hat{\delta}, \hat{p}, \hat{\alpha}, \hat{\beta}, \hat{k}, \hat{s}_1, \ldots, \hat{s}_k, \hat{\lambda}_1, \ldots, \hat{\lambda}_{k+1}),$$



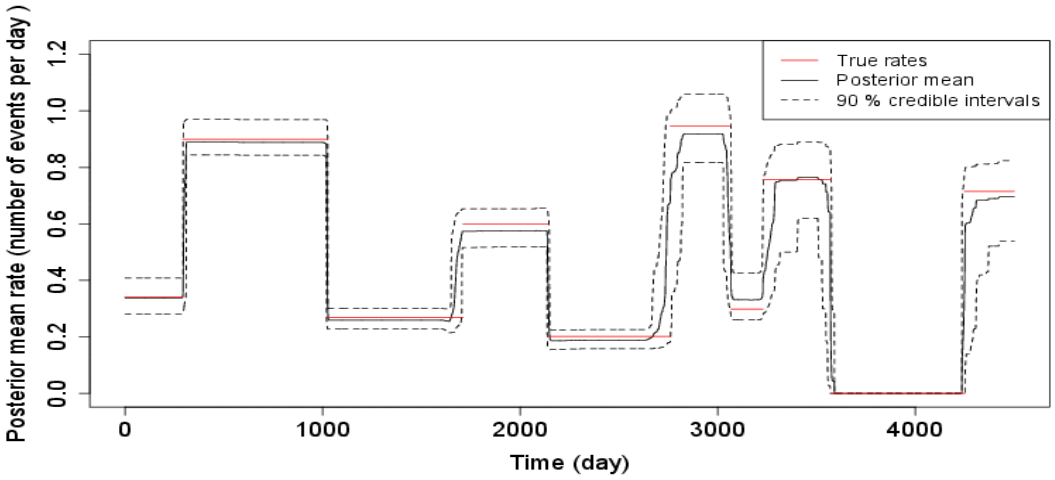
(a) Posterior distribution of k.



(b) Posterior probability of zero rate.



(c) Poster distribution of $\boldsymbol{s}$.



(d) Posterior mean rate.

FIGURE 2.14: Simulation study III, under the penalty prior for rates with $\phi = 2$.

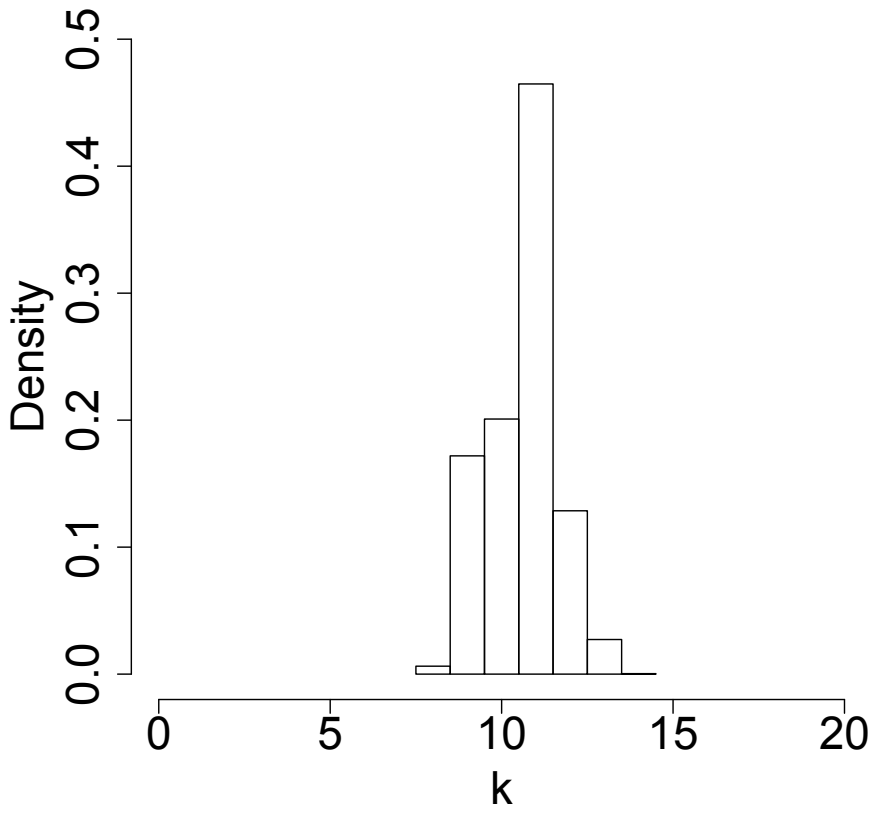


FIGURE 2.15: Real data analysis: Posterior distribution of k, under the mixture prior for rates with $\phi = 2$.

as illustrated in Figure 2.20, we can predict future events $\{\tau_i\}_{i>N}$ according to predictive change points $\{\tilde{s}_j\}_{j>\hat{k}}$ and predictive rates $\{\tilde{\lambda}_j\}_{j>\hat{k}+1}$ shown in Figure 2.21. Specifically, the prediction procedure can be described in four steps:

First, predict future change points $\{\widetilde{s}_j\}$ for $j \geqslant k+1$. As $(\widetilde{s}_{j+1} - \widetilde{s}_j \mid \widehat{\Theta}) \sim \text{Exp}(\hat{\delta})$, one generates $\Delta s_j \overset{\text{iid}}{\sim} \text{Exp}(\hat{\delta})$, and sets $\widetilde{s}_{j+1} = \widetilde{s}_j + \Delta s_j$ with $\tilde{s}_k = \hat{s}_k$.

Second, predict future rates $\{\widetilde{\lambda}_j\}$ for $j \geqslant k+2$ by acceptance-rejection sampling. Particularly, sample $\tilde{z}$ from $\text{Ber}(\hat{p})$. If $\tilde{z} = 0$, set $\tilde{\lambda}_j = 0$; otherwise, the predictive

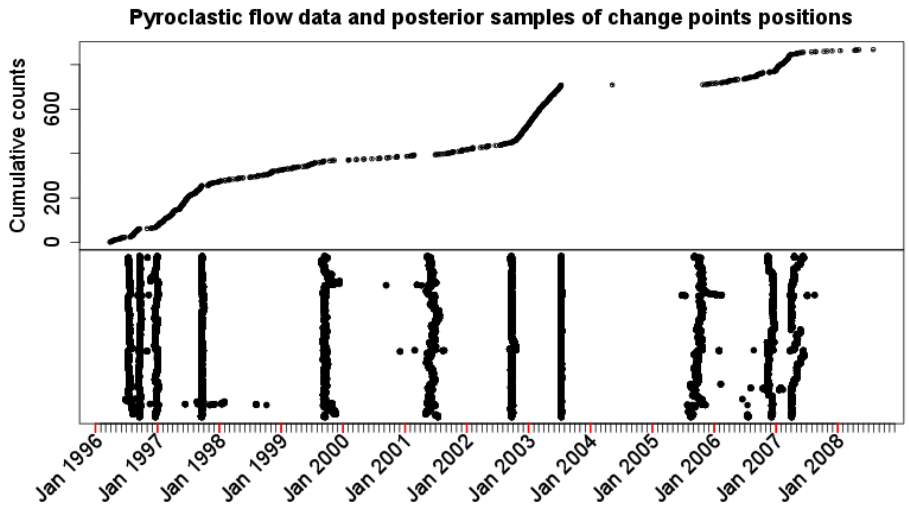


FIGURE 2.16: Real data analysis: Posterior distribution of $\boldsymbol{s}$, under the mixture prior for rates with $\phi = 2$.

density of $\tilde{\lambda}_j$ given the current values of other parameters is:

$$f(\widetilde{\lambda}_j) = \frac{1}{C_1^{(\hat{\alpha},\hat{\beta},\phi)}} \widetilde{\lambda}_j^{\hat{\alpha}-1} e^{-\hat{\beta}\widetilde{\lambda}_j} \mid \widetilde{\lambda}_j - \widehat{\lambda}_{j-1} \mid^{\phi} .$$

Let g denote a mixture of Gamma distributions $\rho\text{Gamma}(\hat{\alpha}, \hat{\beta}) + (1-\rho)\text{Gamma}(\hat{\alpha}+\phi, \hat{\beta})$ and sample $\hat{\lambda}_j$ from g.

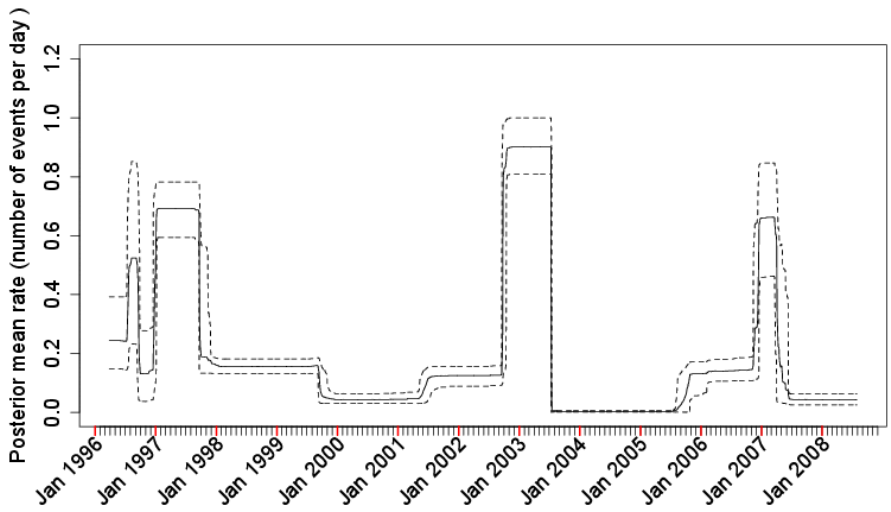


FIGURE 2.17: Real data analysis: Posterior mean rate, under the mixture prior for rates with $\phi = 2$.

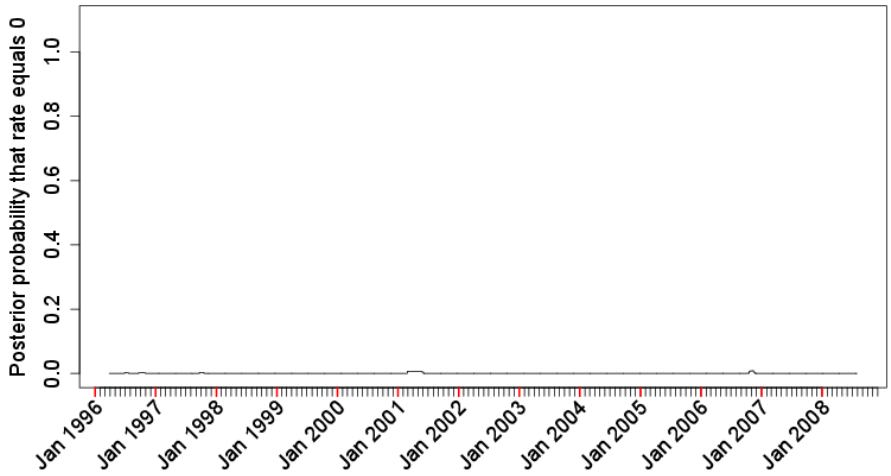


FIGURE 2.18: Real data analysis: Posterior probability or zero rate, under the mixture prior for rates with $\phi = 2$.

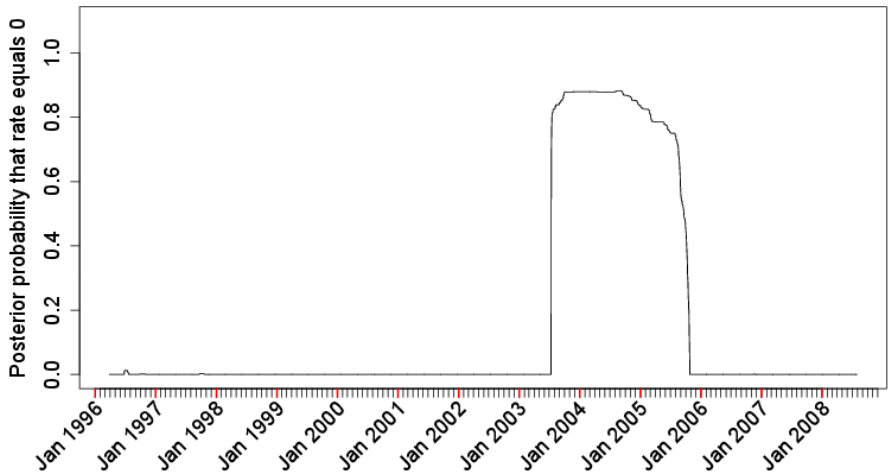


FIGURE 2.19: Real data analysis: Posterior probability, under the mixture prior for rates with $\phi = 2$. (remove the point in the long gap in the pyroclastic flow data)

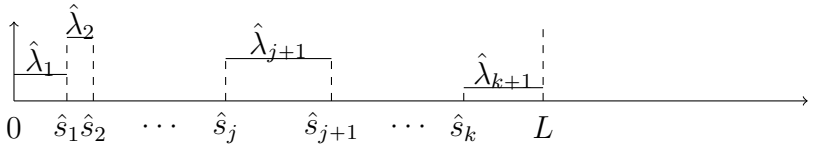


FIGURE 2.20: An illustration of estimates of change points and rates.

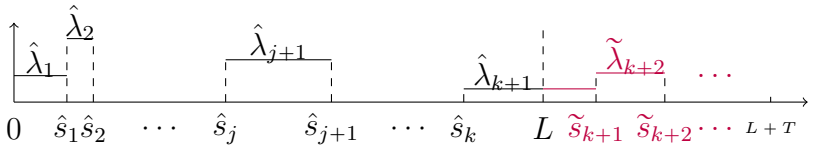


FIGURE 2.21: An illustration of predicted change points and rates.

Accept the new sample with probability f/Mg, with M is the upper limit of f/g. In order to make the acceptance probability as large as possible, the optimal value for ρ is

$$\rho = \frac{\Gamma(\hat{\alpha})\hat{\beta}^{\phi}\widehat{\lambda}_{j-1}^{\phi}}{\Gamma(\hat{\alpha})\hat{\beta}^{\phi}\widehat{\lambda}_{j-1}^{\phi} + \Gamma(\hat{\alpha}+\phi)}.$$

Third, predict future events between $\hat{s}_j$ and $\hat{s}_{j+1}$ by generating $\Delta\tau_n \overset{\text{iid}}{\sim} \text{Exp}(\hat{\lambda}_{j+1})$ and set $\tilde{\tau}_{n+1} = \tilde{\tau}_n + \Delta t_n$. Figure 2.22 shows two sample prediction results. In each of them, the future events are plotted after the observed pyroclastic flows, and future rates are also plotted after the last estimated rate.

Fourth, compute the predictive probability from $\{\tilde{\tau}_{N+1}, \tilde{\tau}_{N+2}, \ldots\}$ after RJMCMC runs. Figure 2.23 presents the probability that at least one large pyroclastic flow will occur in the following 10,000 days (about three years). The curve increases to nearly one in about half a year.

2.6 Conclusion and discussion

In summary, we try to detect multiple change points in the pyroclastic flow dataset. The methods improve upon the RJMCMC method by using special prior distributions to meet particular requirements. Although they induce some difficulty in calculating normalizing constants, the penalized prior is very flexible and the mixture prior is of practical significance. These improvements can be easily adopted in other natural hazard datasets. There are a few ways to extend the work, such as risk assessment, incorporating the predictive probability of pyroclastic flows into the

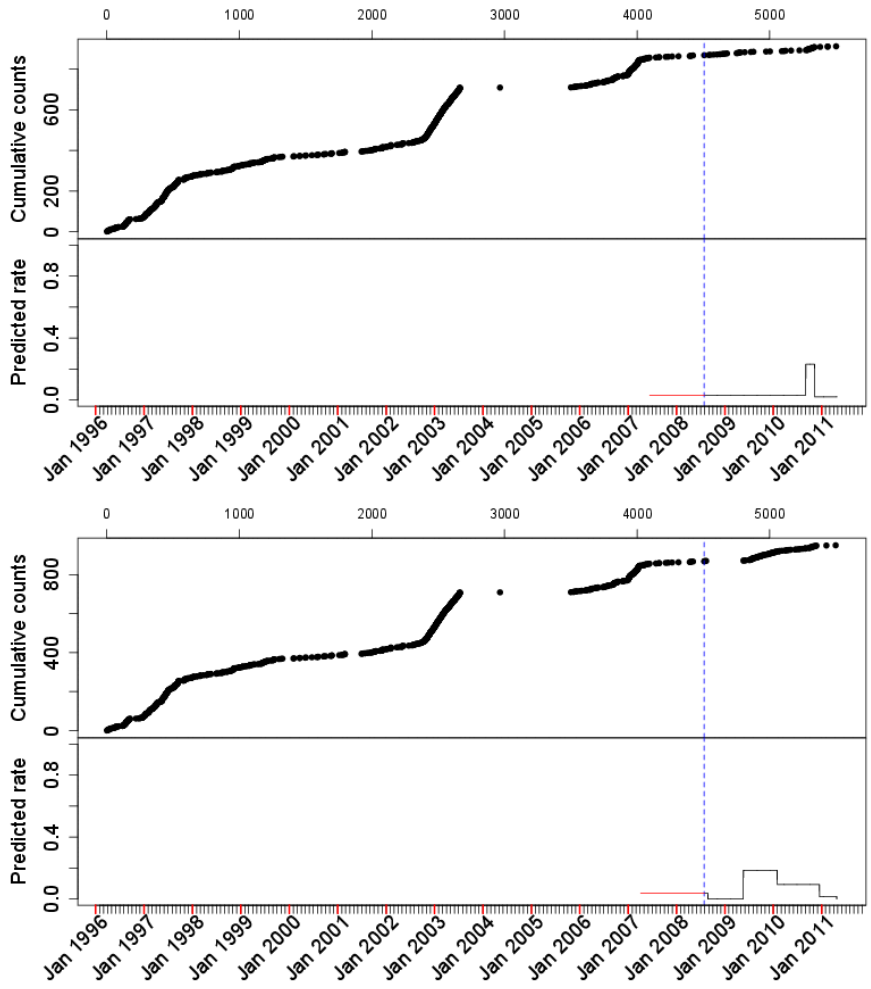


FIGURE 2.22: Two sample predictions: Predictive future rates and pyroclastic flow events.

drawing of hazard map, and relate different volcanic data to explore and establish regression between them.

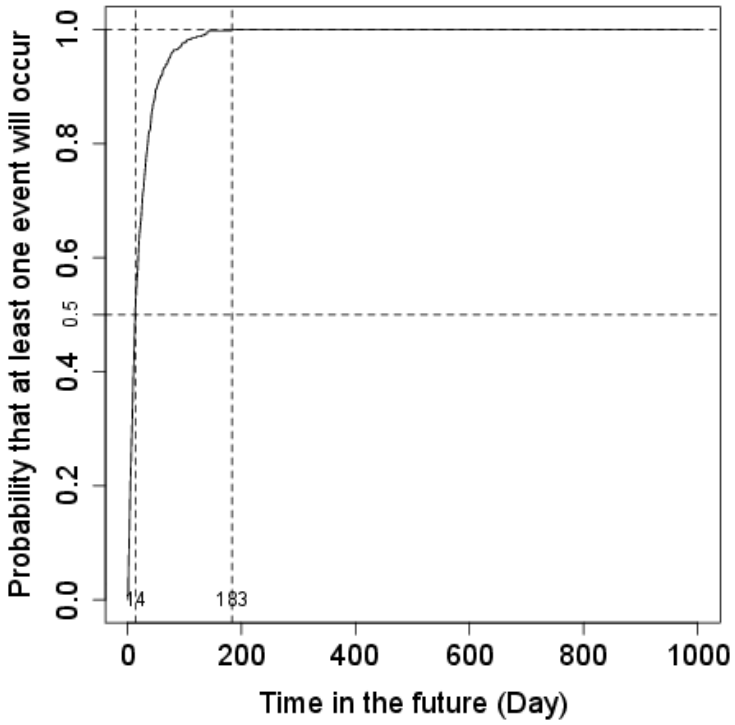


FIGURE 2.23: Predictive probability of occurrence of pyroclastic events in the future.

3

Generalized Regression with a Non-Stationary Markov Process

3.1 Introduction

In Chapter 2, we studied the frequency of pyroclastic flows to predict the probability that a catastrophic event will occur in a certain period of time. In this chapter, we will develop new statistical techniques to model and relate three complex processes and data sets: the process of extrusion of magma into the lava dome which is the cause of pyroclastic flows, the growth of the dome as measured by its height, and the rockfalls which are small rock avalanches off the dome that are an indication of the dome's instability.

Building a joint model for these processes is not only important for understanding the geophysical process, but also to develop methods to predict or estimate important geophysical quantities such as the extrusion rate – key for predicting pyroclastic flows – from easy-to-measure features such as dome height and rockfall frequency.

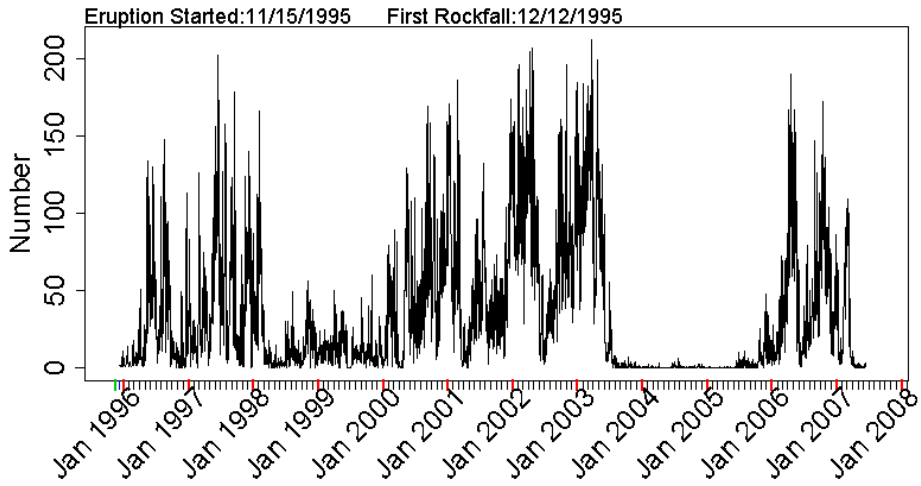


FIGURE 3.1: Daily rockfall counts from December 1995 to June 2007.

3.2 Rockfall data and an initial Negative Binomial model

A rockfall is simply a quantity of rock falling down a cliff or slope. We will be considering rockfalls running down a side of the volcanic dome, and often continuing a modest distance down the mountain. These rockfalls arise from instability in the dome, arising from magma extrusion, weathering (especially rain and temperature changes), and ground tremors.

In this section, we model the distribution of rockfalls. The daily number of rockfalls observed over an approximately 12 year period (12/12/1995–06/13/2007) is plotted as a time series in Figure 3.1. It can be seen that rockfall counts vary significantly over time. The range of the data is from 0 to 212, indicating that the distribution of the daily rockfall count has a much heavier tail than a Poisson distribution; this is clearly seen from Figure 3.2. Another natural possibility to be considered is the Negative Binomial distribution.

The Negative Binomial distribution, denoted by $\mathrm{NB}(\alpha, p)$, is a discrete probability distribution of the number of successes in a sequence of Bernouli trials before a

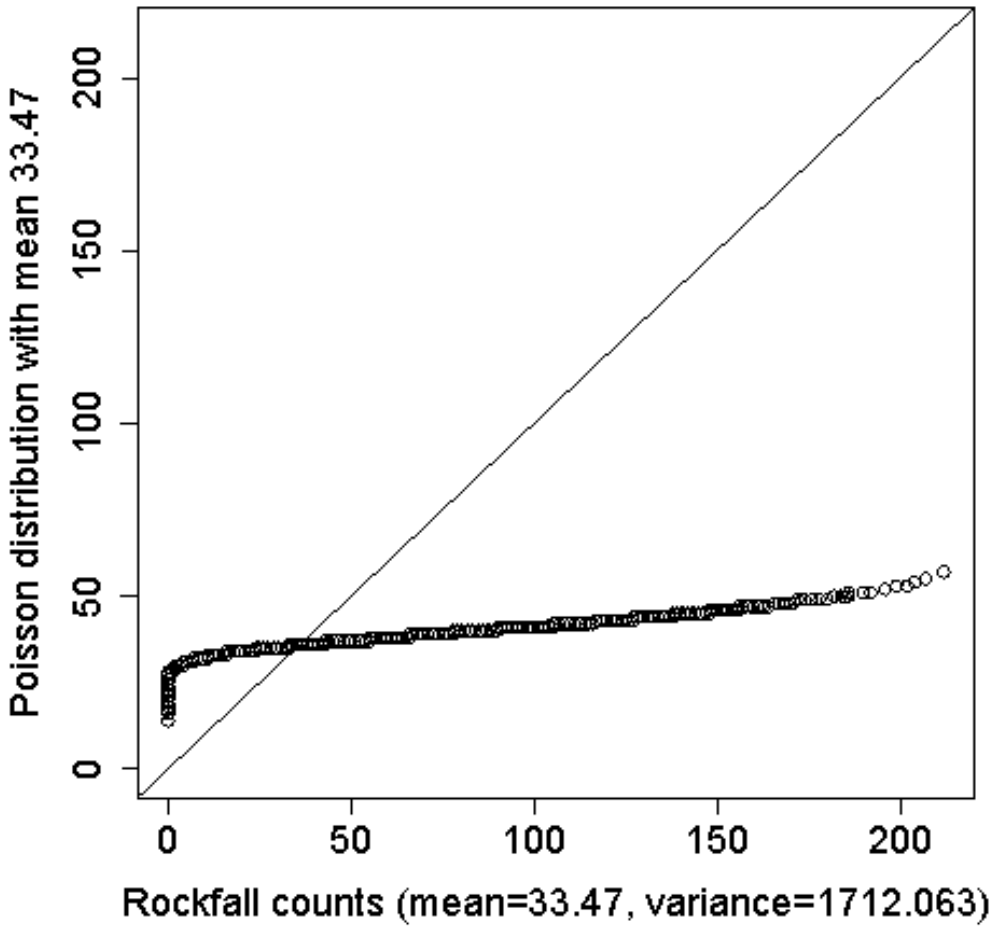


FIGURE 3.2: QQ plot of the rockfall data versus a Poisson distribution.

specified number of failures occur. The probability mass function is

$$P[X = k] = \frac{\Gamma(\alpha + k)}{\Gamma(\alpha)} p^{\alpha} q^{k}, \quad q = 1 - p,$$

where $k \in \{0, 1, 2, \ldots\}$ is the number of successes, $\alpha > 0$ is the number of failures until the experiment is stopped, and q is the success probability in each experiment. It is possible to extend the Negative Binomial distribution to the case where α is a positive real number. In addition, it is often convenient to reparameterize to $\beta = p/q \in (0, \infty)$,

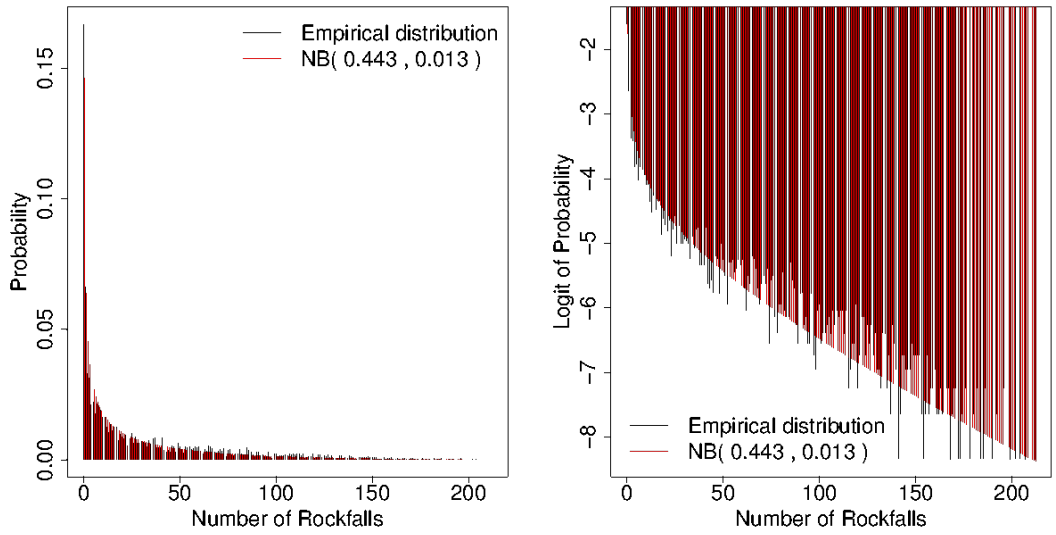


FIGURE 3.3: Comparing the empirical distribution of rockfall data to its maximium likelihood fit to a Negative Binomial distribution NB(α,p).

so that $p^\alpha q^k$ becomes $\beta^\alpha(1+\beta)^{-\alpha-k}$; then the Negative Binomial distribution will be denoted NB(α, β). The mean of NB(α, β) is α/β and the variance is $\alpha(1+\beta)/\beta^2$.

We show, in Figure 3.3, the empirical distribution of rockfall counts together with a Negative Binomial distribution fit to the data using maximum likelihood estimates, $\hat{\alpha} = 0.443$ and $\hat{p} = 0.013$. In the plot on the right, the y-axis of probability is on a logit scale in order to examine the tail more closely. We can see that there are fewer than expected days in which the rockfall counts are quite high, say over 175, which seems to indicate a lack of fit. However, this may well be due to missing data. For instance, 175 rockfalls in one day would mean an average of only 8.2 minutes between rockfalls, and probably much shorter periods between some successive rockfalls, so that different rockfall events could become lumped together. Another possible source of missing data is that activity leading to much more than 175 events might be high enough to have some part of the activity recorded as a small pyroclastic flow.

Figure 3.4 shows a QQ plot and the autocorrelation plot corresponding to the

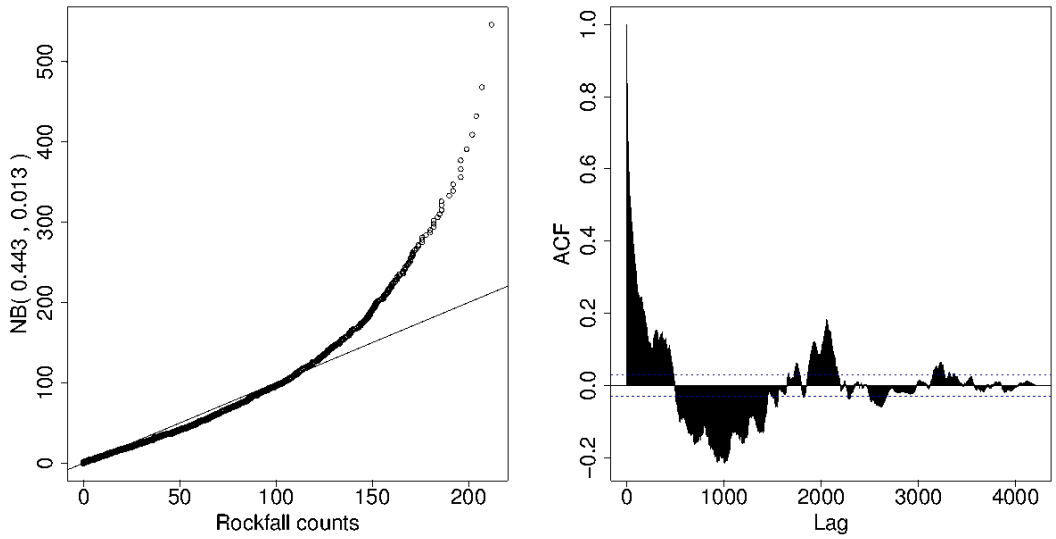


FIGURE 3.4: QQ plot and autocorrelation plot comparing the rockfall data with the maximum likelihood Negative Binomial fit.

maximum likelihood Negative Binomial fit. This tells the same story as the distribution plots, reaffirming that the fit to the negative binomial distribution is quite good for rockfall counts less than 150, yet has an obviously heavier tail than the empirical distribution of rockfalls. In addition to the possibility of having missing data, the autocorrelation plot shows another likely source of the problem: the daily rockfalls seem to be strongly correlated. Hence we turn to a more involved correlated model of rockfall counts.

3.3 Negative Binomial branching process

The principal causes of rockfalls include geology and climate factors, both of which would suggest correlations between days. For instance, a rainy period might lead to a number of high-rockfall days, as might periods of geological instability due to earthquakes. Also, rockfall counts can be expected to differ according to the size of the dome. Hence, we will introduce an AR(1)-like process to model the rockfall data

instead of considering them as a sample independently drawn from the population.

A Negative Binomial branching process Y_t is a stationary, Markov, time-reversible process with the marginal distribution $\mathrm{NB}(\alpha, p)$ and one-step correlation $\mathrm{Corr}(Y_t, Y_{t+1}) = \rho = e^{-\lambda}$. There has not been much work on inference about this process since its intial introduction (Edwards and Gurland, 1961), until a good way to evaluate the likelihood function for an observed dataset $\boldsymbol{y} = \{y_0, \ldots, y_T\}$ of values of some random variables $\{Y_0, \ldots, Y_T\}$ was recently proposed (Wolpert and Brown, 2011). One indirect way is to use data augmentation, which is a recursive update scheme from Y_{t-1} to Y_t. This approach is useful in simulating a Negative Binomial branching process as well as generating the likelihood function for this process. For $1 \leqslant t \leqslant T$, the augmentation scheme is as follows:

$$
\begin{aligned}
&Y_0 \sim \mathrm{NB}\left(\alpha, p\right), \\
&\xi_t \sim \mathrm{Bi}\left(y_{t-1}, \frac{\rho p}{1-\rho+p\rho}\right), \quad \zeta_t \sim \mathrm{NB}\left(\alpha+\xi_t, \frac{\rho}{1-\rho+p\rho}\right), \\
&Y_t = \xi_t + \zeta_t.
\end{aligned}
$$

To calculate the full likelihood, we first obtain conditional probabilities using data augmentation,

$$
\begin{aligned}
P_{ij}(\alpha, p, \rho) &= \mathrm{P}[Y_t = j \mid Y_{t-1} = i] \\
&= \sum_{\xi=0}^{i \wedge j} \binom{i}{\xi} (\rho r)^{\xi} (1-\rho r)^{i-\xi} \frac{\Gamma(\alpha+j)}{\Gamma(\alpha+\xi)(j-\xi)!} r^{\alpha+\xi} (1-r)^{j-\xi},
\end{aligned}
$$

where $r = \dfrac{\rho}{1-\rho+p\rho}$, and then factorize the full likelihood as a series of products

of conditional probabilities:

$$\begin{aligned}
\mathrm{P}[\mathbf{Y}=\mathbf{y}|\alpha,p,\rho] =& \frac{\Gamma(\alpha+y_1)}{\Gamma(\alpha)y_1!}p^\alpha q^{y_1}\prod_{1<t\leqslant n} P_{y_{t-1},y_t}(\alpha,p,\rho) \\
=& \frac{\Gamma(\alpha+y_1)}{\Gamma(\alpha)y_1!}p^\alpha(1-p)^{y_1}\cdot\prod_{1<t\leqslant n}\frac{\Gamma(\alpha+y_t)}{\Gamma(\alpha)y_t!}\cdot\frac{p^\alpha(1-\rho)^{y_t+y_{t-1}}(1-p)^{y_t}}{(1-\rho+\rho p)^{\alpha+y_t+y_{t-1}}} \\
& \cdot\sum_{\xi=0}^{y_j\wedge y_{t-1}}\frac{y_{t-1}!\Gamma(\alpha)y_t!}{(y_{t-1}-\xi)!\Gamma(\alpha+\xi)(y_t-\xi)!\xi!}\left(\frac{\rho}{(1-\rho)^2}\frac{p^2}{(1-p)}\right)^\xi \\
=& \frac{\Gamma(\alpha+y_1)}{\Gamma(\alpha)y_1!}p^\alpha(1-p)^{y_1}\cdot\prod_{1<t\leqslant n}\frac{\Gamma(\alpha+y_t)}{\Gamma(\alpha)y_t!}\cdot\frac{p^\alpha(1-\rho)^{y_t+y_{t-1}}(1-p)^{y_t}}{(1-\rho+\rho p)^{\alpha+y_t+y_{t-1}}} \\
& \cdot {}_2F_1\left(-y_{t-1},-y_t;\alpha;\frac{\rho}{(1-\rho)^2}\frac{p^2}{(1-p)}\right),
\end{aligned}$$

where ${}_2F_1(a,b;c;z)$ is Gauss' hypergeometric function. This leads to a closed form expression for the full likelihood.

3.4 Nonstationarity and model selection

3.4.1 Nonstationary models

Examination of Figure 3.1 suggests another problem: that the process is not stationary. Indeed, the rockfall activity is strongly related to volcanic activity which has "on" and "off" periods and the degree of rockfall activity tends to be very different in each. Even the "on" periods may be very different from each other and the "off" period very different from each other, because of differences in geological features (such as dome height) or climate features during the period.

This nonstationarity will be addressed by utilizing different Negative Binomial Branching models in each rockfall period, i.e., allowing different choices of the three parameters α, p (or β) and ρ. Unfortunately, allowing all three parameters to vary makes the model too complex to work with (especially to address the generalized regression application to be considered later) and also might well overfit the data.

Hence we will search for a lower dimensional parameter-varying model.

In order to gain some insight into possibly appropriate parameter-varying negative binomial branching models for the rockfall data, we first conduct an exploratory study, in which the dataset is divided into 20 time segments, each of will be fit separately to a negative binomial branching model. In Figure 3.5, the red dotted lines are plotted at the change points chosen to divide the rockfall data into the 20 regions. These change points were chosen based on the observation that, in each of the 20 regions, the marginal distribution of the rockfall counts seem to be the same at each time point.

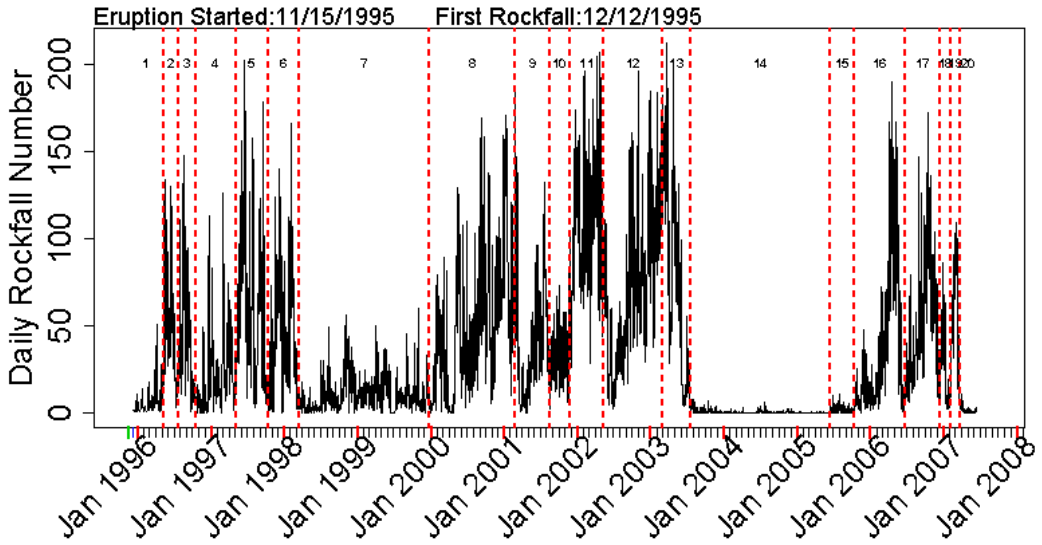


FIGURE 3.5: Rockfall data with the 20 regions to be fitted separately to a negative binomial branching model.

First, to see if we are on the right track by fitting separate models to regions, we first consider fitting simple negative binomial models to each region. Figure 3.4 presents the Q-Q plots of the fits, in each of the 20 regions, of the rockfall data to the maximum likelihood negative binomial model for that region. Generally speaking, the fits are better, especially in the tail regions, so this is a move in a good direction. We know that it is still not correct, however, in that in does not account for the clear

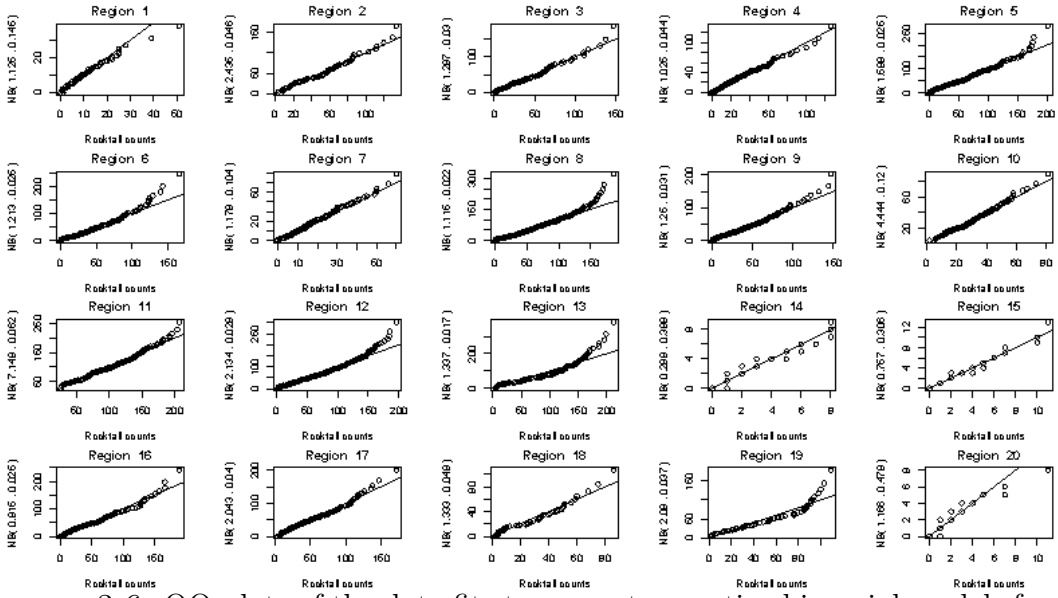


FIGURE 3.6: QQ-plots of the data fits to separate negative binomial models for each of the 20 regions.

dependence in the data, even within a given region. Hence we turn to utilization of utilization of negative binomial branching models with changing parameters.

Let $\text{bNB}(\alpha_i, \beta_i, \rho_i)$ denote the model in each region $i \in \{1, 2, \ldots, 20\}$. Table 3.1 lists the corresponding parameter maximum likelihood estimates for each region. Our goal is to try to find a lower dimensional structure in these values that seems reasonably stable across regions; we can then use the resulting lower dimensional parameter-varying negative binomial branching model as our final model for rockfalls.

Table 3.1: Maximum likelihood estimates of Negative Binomial Branching model parameters for each of the 20 regions.

Region	1	2	3	4	5	6	7	8	9	10
α^{mle}	1.16	2.25	1.21	0.98	1.54	0.98	0.93	0.80	1.21	4.10
β^{mle}	0.17	0.05	0.03	0.04	0.03	0.02	0.09	0.02	0.03	0.12
ρ^{mle}	0.67	0.79	0.81	0.78	0.80	0.80	0.75	0.92	0.89	0.54
Region	11	12	13	14	15	16	17	18	19	20
α^{mle}	6.50	1.77	1.29	0.31	0.76	1.04	2.02	1.49	1.30	1.22
β^{mle}	0.06	0.03	0.95	0.06	0.02	0.02	0.69	0.45	0.03	0.04
ρ^{mle}	0.72	0.91	0.95	0.47	0.33	0.87	0.85	0.82	0.90	0.46

To compare the variability between data sets with different units or widely different means, it is common to use the coefficient of variation. The coefficient of variation (CV) is defined as the ratio of the standard deviation to the mean:

$$c_v = \frac{\sigma}{\mu}.$$

It shows the extent of variability in relation to mean of the population. The coefficient of variation has been commonly used in renewal theory, queueing theory, and reliability theory. In these fields, the exponential distribution is often more important than the normal distribution. The standard deviation of an exponential distribution is equal to its mean, so its coefficient of variation is equal to 1. Therefore, in these applied probability fields, distributions with CV < 1 are considered low-variance, while those with CV > 1 are considered high-variance.

The empirical coefficients of variation (i.e, the sample standard deviation divided by the sample mean) for each of the three lists of MLEs in Table 3.1 are:

$$c_v(\alpha^{mle}) = 0.84, \quad c_v(\beta^{mle}) = 1.7, \quad c_v(\rho^{mle}) = 0.23.$$

This suggests that the bNB parameters α and β are more variable than ρ, and suggests the possibility of assuming that ρ is constant across regions. This is especially plausible when realizing that only regions 14, 15 and 20 had values that significantly differed from the others, and these were regions of very low rockfall activity so that the likelihoods were relatively flat in these regions; the resulting mle's are thus not particularly reliable. So we will, henceforth, assume a constant (but unknown) ρ in the negative binomial branching process.

A further possible simplification is to find a one dimensional function of α and β that is constant. We will consider three such models:

- Model 1: β is constant.

- Model 2: α is constant.
- Model 3: $C_{\alpha\beta} = \alpha\beta$ is constant; this is a compromise between Model 1 and Model 2.

Note that, for each model (and assuming constant ρ) there is still one degree of freedom left in choice of the negative binomial branching models for each region. For analysis, it is convenient to reparameterize so that this free parameter is the mean of the negative binomial branching model in that region, namely $\mu_i = \alpha_i/\beta_i$. Hence the three models to be considered have the following unknown parameters (generalizing to the scenario of k change points):

- $M_1 : \theta_1 = \{\mu_1, \ldots, \mu_{k+1}, \beta, \rho\}$.
- $M_2 : \theta_2 = \{\mu_1, \ldots, \mu_{k+1}, \alpha, \rho\}$.
- $M_3 : \theta_3 = \{\mu_1, \ldots, \mu_{k+1}, C_{\alpha\beta}, \rho\}$.

These three models are not nested, but the number of parameters is the same: $k+3$. Each model has a distinct parameter, and the other $k+2$ parameters are same, simplifying the model selection problem

3.4.2 Prior distributions and likelihood

It is reasonable to assign equal probabilities to the three models: $\pi_1(M_1) = \pi_2(M_2) = \pi_3(M_3) = 1/3$. Furthermore, we will use a prior distribution for $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_{k+1}\}$ and ρ under which $\boldsymbol{\mu}$ and ρ are independent. This independence makes sense because our prior information about these two parameters arises from different sources. The level of $\boldsymbol{\mu}$ largely depends on the volcanic activity, while the correlation is more related to climate or other geological factors. Since $\boldsymbol{\mu}$ and ρ appear in all three

models, we can assign improper priors for $\boldsymbol{\mu}$ and a uniform prior for ρ:

$$\pi_j(\mu_i) \propto \frac{1}{\mu_i}, \qquad j = 1, 2, 3, \quad i = 1, \ldots, k+1,$$

$$\pi_j(\rho) = 1, \qquad j = 1, 2, 3.$$

For the distinct parameters in the three models, β in M_1, α in M_2 and $C_{\alpha\beta}$ in M_3, it is difficult to specify compatible priors, as the parameters have very difference effects on the likelihood. We will instead simply present the marginal likelihoods as a function of these parameters, evaluated at a number of values.

In each of the twenty regions, the process is assumed to be stationary, so the likelihood function discussed in Section 3.3 applies. One first needs to reparameterize each likelihood according to the selected model – e.g., for Model 1, just replace α and p in $\text{bNB}(\alpha, p, \rho)$ by $p = \dfrac{\beta}{1+\beta}$ and $\alpha = \mu\beta$); we will abuse notation and write the ensuing model as $\text{bNB}(\boldsymbol{\mu}, \beta, \rho)$.

3.4.3 Marginal likelihoods

Model selection will be done by looking at the marginal likelihoods corresponding to each model (at various values of the fixed parameters). One obtains the marginal likelihoods simply by multiplying the (reparameterized) likelihoods by the priors in the previous section, and integrating out over the parameters. It is shown in Appendix C that these marginal likelihoods are finite. The expressions for these integrals, at fixed values of the distinct parameters (denoted by α^*, β^* and $C^*_{\alpha\beta}$), are as follows:

$$
\begin{aligned}
m_1(\boldsymbol{x})_{\beta^*} &= \int L(\boldsymbol{x} \mid \mu_1, \ldots, \mu_{k+1}, \beta = \beta^*, \rho)\pi_1(\mu_1, \ldots, \mu_{k+1}, \rho)\, d\mu_1 \cdots d\mu_{k+1}\, d\rho \\
&= \int \left[\prod_{i=1}^{k+1} \mathrm{NB}(\boldsymbol{x} \mid \mu_i, \rho, \beta = \beta^*)\right] \pi_1(\rho)\pi_1(\mu_1) \cdots \pi_1(\mu_{k+1})\, d\mu_1 \cdots d\mu_{k+1}\, d\rho \\
&= \int_\rho \prod_{i=1}^{k+1} \left[\int_{\boldsymbol{\mu}} \mathrm{NB}(\boldsymbol{x} \mid \mu_i, \rho, \beta = \beta^*)\pi_1(\mu_i)\, d\mu_i\right] \pi_1(\rho)\, d\rho, \\
m_2(\boldsymbol{x})_{\alpha^*} &= \int L(\boldsymbol{x} \mid \mu_1, \ldots, \mu_{k+1}, \alpha = \alpha^*, \rho)\pi_2(\mu_1, \ldots, \mu_{k+1}, \rho)\, d\mu_1 \cdots d\mu_{k+1}\, d\rho \\
&= \int_\rho \prod_{i=1}^{k+1} \left[\int_{\boldsymbol{\mu}} \mathrm{NB}(\boldsymbol{x} \mid \mu_i, \rho, \alpha = \alpha^*)\pi_2(\mu_i)\, d\mu_i\right] \pi_2(\rho)\, d\rho, \\
m_3(\boldsymbol{x})_{C^*_{\alpha\beta}} &= \int L(\boldsymbol{x} \mid \mu_2, \ldots, \mu_{k+1}, C_{\alpha\beta} = C^*_{\alpha\beta}, \rho)\pi_3(\mu_1, \ldots, \mu_{k+1}, \rho)\, d\mu_2 \cdots d\mu_{k+1}\, d\rho \\
&= \int_\rho \prod_{i=1}^{k+1} \left[\int_{\boldsymbol{\mu}} \mathrm{NB}(\boldsymbol{x} \mid \mu_i, \rho, C_{\alpha\beta} = C^*_{\alpha\beta})\pi_3(\mu_i)\, d\mu_i\right] \pi_3(\rho)\, d\rho,
\end{aligned}
$$

where $\mathrm{NB}(\boldsymbol{x} \mid \mu_i, \rho_i, \cdot)$ is the likelihood function with data in the i-th region.

3.4.4 Approximating the integrals

The marginal likelihood functions for the different models can be approximated in similar ways, as the integrals have similar forms. In the following, we will present the computation of $m_1(x)_{\beta^*}$ step by step using Laplace's method (Laplace, 1974).

(a) In each of the 20 regions, there are hundreds of data, which makes the integrals $\int_{\mu_i} \mathrm{NB}(\boldsymbol{x} \mid \mu_i, \rho, \beta = \beta^*)\pi_1(\mu_i)\, d\mu_i$ $(i = 1, \ldots, 20))$ very small. When we then numerically integrate ρ, the integrand is essentially zero in statistical software, such as R. Hence, we compute a similar integral:

$$
C_i = \log \int_0^\infty \int_0^\infty \int_0^\infty \mathrm{NB}(\boldsymbol{x} \mid \mu_i, \lambda, \beta)\pi(\mu_i)\pi(\lambda)\pi(\beta)\, d\beta\, d\lambda\, d\mu_i,
$$

for each region and take the constant factor C^* out for computational efficiency and accuracy. This results in the following decomposition of $m_1(\boldsymbol{x})_{\beta^*}$:

$$
\begin{aligned}
m_1(\boldsymbol{x})_{\beta^*} &= \int_\rho \prod_{i=1}^{k+1} \left[\int_{\mu_i} \mathrm{NB}(\boldsymbol{x} \mid \mu_i, \rho, \beta = \beta^*) \pi_1(\mu_i)\, d\mu_i \right] \pi_1(\rho)\, d\rho \\
&= \int_\rho \exp\left\{ \sum_{i=1}^{k+1} \log \int_{\mu_i} \mathrm{NB}(\boldsymbol{x} \mid \mu_i, \rho, \beta = \beta^*) \pi_1(\mu_i)\, d\mu_i \right\} \pi_1(\rho)\, d\rho \\
&= \exp\left(\sum_{i=1}^{k+1} C_i \right) \\
&\quad \cdot \int_\rho \exp\left\{ \sum_{i=1}^{k+1} \left(\log \int_{\mu_i} \mathrm{NB}(\boldsymbol{x} \mid \mu_i, \rho, \beta = \beta^*) \pi_1(\mu_i)\, d\mu_i - C_i \right) \right\} \pi_1(\rho)\, d\rho \\
&= C^* \cdot \int_\rho h(\rho \mid \boldsymbol{x}) \pi_1(\rho)\, d\rho,
\end{aligned}
$$

where $C^* = \exp\left(\sum_{i=1}^{k+1} C_i \right)$ and

$$
h(\rho \mid \boldsymbol{x}) = \exp\left\{ \sum_{i=1}^{k+1} \left(\log \int_{\mu_i} \mathrm{NB}(\boldsymbol{x} \mid \mu_i, \rho, \beta = \beta^*) \pi_1(\mu_i)\, d\mu_i - C_i \right) \right\}.
$$

(b) Compute $\log \int_{\mu_i} \mathrm{NB}(\boldsymbol{x} \mid \mu_i, \rho, \beta = \beta^*) \pi_1(\mu_i)\, d\mu_i$ and C_i using Laplace's method.

To compute C_i, $i = 1, \ldots, k+1$, we first extend the range of the integration out to infinity by changes of variables:

$$
\begin{aligned}
&\int_0^\infty \int_0^\infty \int_0^\infty \mathrm{NB}(\boldsymbol{x} \mid \mu_i, \lambda, \beta) \pi(\mu_i) \pi(\lambda) \pi(\beta)\, d\beta\, d\lambda\, d\mu_i \\
&= \int_{-\infty}^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty \mathrm{NB}(\boldsymbol{x} \mid e^{l\mu_i}, e^{l\lambda}, e^{l\beta}) e^{l\mu_i} e^{l\beta} e^{l\lambda} \pi(e^{l\mu_i}) \pi(e^{l\lambda}) \pi(e^{l\beta})\, dl\beta\, dl\lambda\, dl\mu_i,
\end{aligned}
$$

where $l\mu_i = \log(\mu)$, $l\beta = \log(\beta)$, $l\lambda = \log(\lambda)$.

We can either use diffuse proper priors as

$$\mu_i \sim \text{LN}(m_\mu, \sigma_\mu^2), \quad \beta \sim \text{LN}(m_\beta, \sigma_\beta^2), \quad \lambda \sim \text{LN}(m_\lambda, \sigma_\lambda^2),$$

with $m_\mu = m_\beta = m_\lambda = 0$, $\sigma_\mu = \sigma_\beta = \sigma_\lambda = 10$, which simplies the integral as

$$\int_{\mathcal{R}^3} \text{NB}(\boldsymbol{x} \mid e^{l\mu_i}, e^{l\lambda}, e^{l\beta}) N(l\mu_i \mid m_\mu, \sigma_\mu^2) N(l\lambda \mid m_\lambda, \sigma_\lambda^2) N(l\beta \mid m_\beta, \sigma_\beta^2)\, dl\beta\, dl\lambda\, dl\mu_i;$$

or assign the same prior distributions to μ and λ as those for the models, and a Gamma prior for β:

$$\pi(\mu_i) \propto \frac{1}{\mu_i}, \quad \beta \sim \text{Gamma}(\mu_\beta, \sigma_\beta^2), \quad \lambda \sim \text{Exp}(1),$$

with $\mu_\beta = 0.1, \sigma_\beta^2 = 0.06$ which were picked based on the information about MLEs. The results of C_i $(i = 1, \ldots, 20)$ under different prior distributions are listed in Table 3.2. It is straightforward to observe that all the values are small and the pairs for different priors do not differ much.

The approximation method for the integrals $\log \int_{\mu_i} \text{NB}(\boldsymbol{x} \mid \mu_i, \lambda, \beta = \beta^*) \pi_1(\mu_i)\, d\mu_i$ and $\int_\rho h(\rho \mid \boldsymbol{x}) \pi_1(\rho)\, d\rho$ is similar to the above approach.

3.4.5 Comparing the marginal likelihoods of the three proposed models

Tables 3.3-3.5, give the marginal likelihood of each model for various values of the distinct parameter of each model. Figure 3.7 compares the logarithm of the marginal likelihoods of the three models. From the results, it is clear that Model 3 has by far the largest marginal likelihood, and hence will be the model we ultimately use for the rockfalls.

3.5 Correlated Negative Binomial regression

The ongoing eruption of the Soufrière Hills Volcano (SHV) on Montserrat involves lava extrusion, lava dome growth, dome collapses and pyroclastic flows. In order

Table 3.2: Values of C_i in 20 regions.

i	Log-normal prior	Improper and Gamma priors
1	-410.380	-404.196
2	-335.670	-329.761
3	-371.635	-365.881
4	-760.451	-754.536
5	-773.843	-768.029
6	-683.103	-677.324
7	-2020.034	-2014.012
8	-1834.424	-1829.475
9	-686.898	-681.649
10	-414.066	-407.609
11	-826.006	-819.822
12	-1314.771	-1309.678
13	-589.162	-584.661
14	-569.854	-564.060
15	-219.510	-213.358
16	-993.481	-988.038
17	-757.127	-751.512
18	-201.939	-196.182
19	-213.454	-208.212
20	-127.998	-122.592

Table 3.3: Marginal likelihood for Model 1.

β^*	$m_1(x)/e^{-13990.59}$
0.010	2.950×10^{-52}
0.020	7.923×10^{-50}
0.030	1.458×10^{-46}
0.035	$\mathbf{3.635\times10^{-46}}$
0.040	2.441×10^{-46}
0.045	1.338×10^{-47}
0.046	7.893×10^{-48}
0.048	6.517×10^{-49}
0.050	5.194×10^{-50}

Table 3.4: Marginal likelihood for Model 2.

α^*	$m_2(x)/e^{-13990.59}$
0.20	4.842×10^{-95}
0.30	1.022×10^{-68}
0.40	1.657×10^{-53}
0.50	1.270×10^{-44}
0.60	4.878×10^{-40}
0.65	5.597×10^{-39}
0.70	$\mathbf{1.412\times10^{-38}}$
0.75	8.648×10^{-39}
0.80	1.861×10^{-39}
0.90	4.018×10^{-42}
1.00	3.138×10^{-46}

Table 3.5: Marginal likelihood for Model 3.

$C^*_{\alpha\beta}$	$m_3(x)/e^{-13990.59}$
0.010	1.344×10^{-16}
0.020	1.249×10^{-08}
0.025	3.016×10^{-07}
0.026	4.299×10^{-07}
0.028	7.373×10^{-07}
0.030	9.047×10^{-07}
0.032	$\mathbf{9.470\times10^{-07}}$
0.034	7.780×10^{-07}
0.035	6.157×10^{-07}
0.040	1.417×10^{-07}
0.045	1.322×10^{-08}
0.050	5.854×10^{-10}
0.070	1.651×10^{-17}
0.090	2.415×10^{-27}

to effectively assess future volcanic hazards at SHV, particularly pyroclastic flows, monitoring extrusion rates is essential. However, extruded lava cannot be observed directly, but can only be estimated.

For the most part, extruded lava translates directly into increase of the volume of the lava dome. It is often difficult, however, to determine the volume of the dome, whereas dome height is straightforward to obtain. Increased lava extrusion

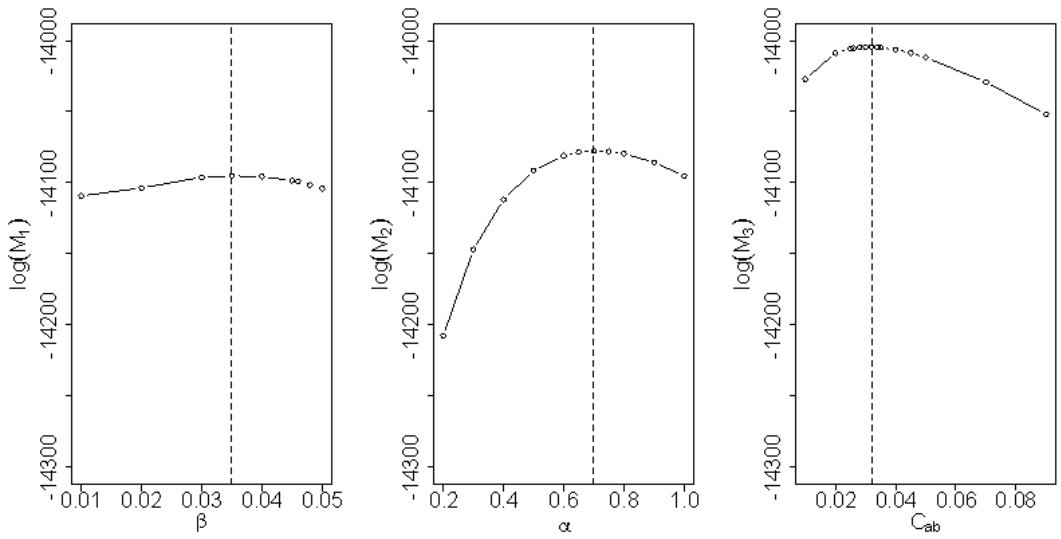


FIGURE 3.7: Log of the marginal likelihoods for the three models.

also seems to translate into increased instability of the dome, and, hence, increased numbers of rockfalls. Thus, if we could determine the relationship between, extrusion rate, dome height and daily rockfall counts, we would be able to solve the inverse problem of predicting extrusion rate based on observed rockfalls and dome height. In this section, we develop a generalized linear regression model relating these three variables.

Our outcome variable will be the number of rockfalls, Y_t, on day t. The measurement error model we utilize for rockfalls is Model 3 from the previous section, namely the non-stationary Negative Binomial branching process with varying parameters, which was denoted bNB$(\mu_t, C_{\alpha\beta}, \rho)$ with $\mu_t = \alpha_t/\beta_t$ being the mean of the process. The true mean $\mu(t)$ is expected to relate to the extrusion rate and dome height, and we utilize log-linear regression to model this relationship.

Luckily, estimates of the extrusion rate are available to assist in developing the regression model. The extrusion rate is the total dense rock equivalent (DRE) volume of extruded andesite magma. The total cumulative lava extrusion is calculated as

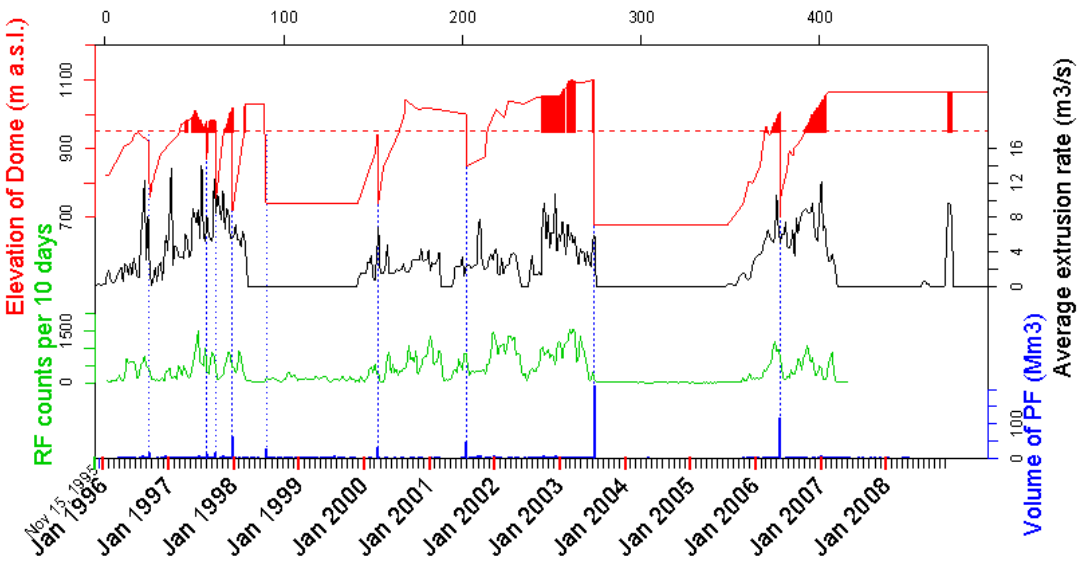


FIGURE 3.8: Time series of four volcanic data processes.

the sum of the change in the lava dome volume, pyroclastic flow deposit (the volume of which was estimated from field measurements) and ash fall deposit volumes, all converted to DRE. The average extrusion rate for every 10 day period was calculated.

Figure 3.8 illustrates four processes, the volume of pyroclastic flows in Mm^3, rockfall counts aggregated for every 10 days (for comparison), average extrusion rate in m^3/s, and elevation of the dome in meters above sea level. For dome height, the dataset contains observations for 86 days and we use linear interpolation for other dates The red regions indicate the days when the dome height was above 950 meters and, simultaneously, the extrusion rate was higher than $4.5m^3/s$; such days were considered particularly "dangerous" periods, since the dome was not stable. The general trends of these three processes are in accordance with each other. The dates when large pyroclastic flows (of volume greater than $9Mm^3$) occurred are highlighted by blue dotted lines, and correspond to those major lava dome collapses.

3.5.1 Model

Let $\{y_t\}_{t=1}^T$ denote the time series of total rockfall counts for every 10 days, and assume it follows a non-stationary branching Negative Binomial process with a mean process $\{\mu_t\}$, and two constant parameters $C_{\alpha\beta} = \alpha_t\beta_t$ and ρ. In addition, we assume that the mean process depends on the independent variables through the log link function. The correlated negative binomial regression model is formulated as below:

$$y_t \sim \text{bNB}(\mu_t, C_{\alpha\beta}, \lambda), \quad t = 1, \ldots, T,$$
$$\log(\mu_t) = \gamma_0 + \gamma_1 \times x_{t1} + \gamma_2 \times x_{t2},$$

where $\text{Corr}[y_s, y_t] = e^{-\lambda|s-t|} = \rho^{|s-t|}$, $\boldsymbol{x_1}$ = log(Extrusion Rate + 1), and $\boldsymbol{x_2}$ = log(Dome Height/average(Dome Height)).

The use of the log linear model is because there should be some power relationship between rockfalls and the dome height and extrusion rate, but it is not at all clear what powers. Adding 1 to the extrusion rate before taking the log simply ensures that zeroes remain zero. In order to make the covariates comparable in values, we first standardize the data by dividing the average dome height; this is primarily to hopefully reduce correlation in the MCMC analysis that will ultimately be needed for the inverse problem.

3.5.2 Prior distribution and likelihood function

The only effect that the non-stationarity has on the likelihood expression of the stationary Negative Binomial branching process is that the appreance of α and/or β is replaced by α_t and/or β_t. The likelihood function for $\text{bNB}(\boldsymbol{\alpha}, \boldsymbol{p}, \rho)$ has the following form, and using the connection between different parametrization: $\boldsymbol{p} = \dfrac{\boldsymbol{\beta}}{1+\boldsymbol{\beta}}$, $\boldsymbol{\alpha} = \sqrt{C_{\alpha\beta}\boldsymbol{\mu}}$, $\boldsymbol{\beta} = \sqrt{\dfrac{C_{\alpha\beta}}{\boldsymbol{\mu}}}$, we can obtain the likelihood function for $\text{bNB}(\boldsymbol{\mu}, C_{\alpha\beta}, \rho)$:

$$\mathrm{P}[\boldsymbol{Y}=\boldsymbol{y}\mid\boldsymbol{\alpha},\boldsymbol{p},\rho]=p(y_1)\prod_{1<j\leqslant T}p(y_j|y_{j-1})$$

$$=\frac{\Gamma(\alpha_1+y_1)}{\Gamma(\alpha_1)y_1!}p_1^{\alpha_1}(1-p_1)^{y_1}\cdot\prod_{1<j\leqslant n}\frac{\Gamma(\alpha_j+y_j)}{\Gamma(\alpha_j)y_j!}\cdot\frac{p_j^{\alpha_j}(1-\rho)^{y_j+y_{j-1}}(1-p_j)^{y_j}}{(1-\rho+\rho p_j)^{\alpha_j+y_j+y_{j-1}}}$$

$$\cdot\sum_{\xi=0}^{y_j\wedge y_{j-1}}\frac{y_{j-1}!\Gamma(\alpha_j)y_j!}{(y_{j-1}-\xi)!\Gamma(\alpha_j+\xi)(y_j-\xi)!\xi!}\left(\frac{\rho}{(1-\rho)^2}\frac{p_j^2}{(1-p_j)}\right)^{\xi}$$

$$=\frac{\Gamma(\alpha_1+y_1)}{\Gamma(\alpha_1)y_1!}p_j^{\alpha_1}(1-p_j)^{y_1}\cdot\prod_{1<j\leqslant T}\frac{\Gamma(\alpha_j+y_j)}{\Gamma(\alpha_j)y_j!}\cdot\frac{p_j^{\alpha_j}(1-\rho)^{y_j+y_{j-1}}(1-p_j)^{y_j}}{(1-\rho+\rho p_j)^{\alpha_j+y_j+y_{j-1}}}$$

$$\cdot\,{}_2F_1(-y_{j-1},-y_j;\alpha_j;\frac{\rho}{(1-\rho)^2}\frac{p_j^2}{(1-p_j)}),$$

where ${}_2F_1(a,b;c;z)$ is Gauss' hypergeometric function.

Finding good objective priors for the parameters $\boldsymbol{\gamma}$, ρ and $C_{\alpha\beta}$ is very difficult so the following vague proper priors will be used, together with sensitivity studies:

$$\gamma_i\sim N(0,\sigma^2),\quad i=0,1,2$$

$$\rho\sim\mathrm{Unif}(0,1);$$

$$C_{\alpha\beta}\sim\mathrm{Gamma}(a,b).$$

The sensitivity analysis involved choices of several sets of prior hyperparameters, such as 10^3 and 10^6 for σ^2, and $(10^{-5},10^{-4})$ and $(10^{-6},10^{-6})$ for (a,b) (so that the prior mean and variance of $C_{\alpha\beta}$ are 0.1 and 10^3, 1 and 10^6 respectively); the results did not show much difference.

Under the proper prior distributions, it is straightforward to check the posterior distributions are also proper, using an analysis similar to Appendix C.

3.5.3 Simulated data

In this section, we develop simulation methods needed for dealing with the dynamic negative binomial branching process. In the process of studying the process, we

first developed a simulation method for bNB(α_t, β, ρ), where parameter α is varying while the other parameters β and ρ remain constant. Refer to Appendix A for details. Although this process is not applied in our analysis to the regression model, the methodology can be very useful in other relevant research areas.

To find a simulation method for the dynamic branching negative binomial process in which both α and β are changing over the time, we turn to another interpretation of the process bNB(α, β, ρ). The process can be viewed as a linear birth/death process with immigration rates:

$$b = \frac{\lambda}{\beta} \quad \text{(Birth rate)},$$

$$d = \frac{\lambda}{p} = \lambda + b \quad \text{(Death rate)},$$

$$i = \lambda\frac{\alpha}{\beta} = \lambda\mu \quad \text{(Immigration rate)},$$

where $p = \dfrac{\beta}{1+\beta}$.

An integer-valued Markov process $\{X_t\}$ can be constructed as a step function with initial values $t_0 = 0$ and $X_0 = x_0$, and for $n \geqslant 0$:

$$\{\tau_n\} \overset{iid}{\sim} \text{Exp}(1),$$

$$t_{n+1} = t_n + \frac{\tau_n}{i + x_n(b+d)},$$

$$X_{n+1} = x_{n+1},$$

$$x_{n+1} = \begin{cases} x_n - 1 \text{ with probability } \frac{x_n d}{i+x_n(b+d)}, \\ x_n + 1 \text{ otherwise.} \end{cases}$$

When the mean process μ_t is a specified non-constant function with derivative μ_t', with constant $\lambda_t = \lambda$ and $\alpha_t\beta_t = C_{\alpha\beta}$, the above construction will need some modification. First of all, $\dfrac{\alpha_t}{\beta_t}$ is not equal to μ_t, but they have the following relation:

$$\alpha_t = \sqrt{C_{\alpha\beta}} \left\{ \mu_t + \frac{\mu_t'}{\lambda} \right\}^{1/2}, \quad \beta_t = \sqrt{C_{\alpha\beta}} \left\{ \mu_t + \frac{\mu_t'}{\lambda} \right\}^{-1/2},$$

or equivalently,

$$i_t = \lambda\mu_t + \mu_t', \quad b_t = \sqrt{\frac{\lambda i_t}{C_{\alpha\beta}}}, \quad d_t = \lambda + b_t.$$

In this case, an integer-valued Markov process $\{X_t\}$ can be constructed as a step function with initial values $t_0 = 0$ and $X_0 = x_0$, and for $n \geqslant 0$:

$$\{\tau_n\} \overset{iid}{\sim} \text{Exp}(1),$$

$$t_{n+1} = \inf \left\{ t > t_n : \int_{t_n}^{t} [i_s + (b_s + d_s)x_n]\, ds > \tau_n \right\},$$

$$X_{n+1} = x_{n+1}, t_n \leqslant t < t_{n+1},$$

$$x_{n+1} = \begin{cases} x_n - 1 \text{ with probability } \frac{x_n d_{t_{n+1}}}{i_{t_{n+1}} + x_n(b_{t_{n+1}} + d_{t_{n+1}})}, \\ x_n + 1 \text{ otherwise.} \end{cases}$$

When rockfall counts drop quickly, such that $i_t = \lambda\mu_t + \mu_t' < 0$, it's not reasonable to impose the modeling condition that the autocorrelation stays constant. To deal with this, let

$$\lambda_t = \max(\lambda, -2\mu_t'/\mu_t)$$

to ensure $i_t = \lambda\mu_t + \mu_t'$ is always positive.

3.5.4 Simulation study

Before using the generalized regression model on the real data, we would like to conduct an experiment on a simulated dataset and see how the model works. We simulate rockfall data according to the correlated negative binomial regression model using the simulation method for the dynamic branching negative binomial process.

In order to generate a dataset that has some features similar to the real rockfall data, we incorporate some information of the real extrusion rate and dome height into the regression covariates. Let

$$\boldsymbol{x_2} = \log(\text{Dome Height}/\text{average}(\text{Dome Height}))$$

denote dome height data, where "Dome height" is the red curve plotted in Figure 3.8. However, for $\boldsymbol{x_1}$, we use a step function of extrusion rate instead of the real data, as illustrated by the blue lines in Figure 3.9. This assumption is reasonable because, in reality, the large jumps of the extrusion rate indicate different phases of dome growth and, because the extrusion rates in the data were only inferred, the smaller fluctuations in the extrusion data are unlikely to be accurate.

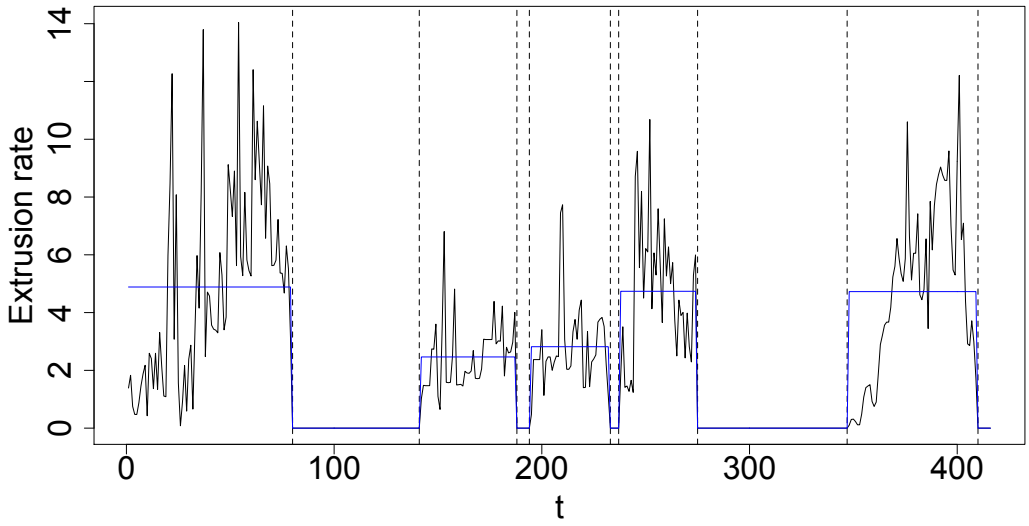


FIGURE 3.9: Step extrusion rate assumed in simulation.

In Figure 3.9, the black curve represents the "real" calculated extrusion rate, from which we recognize several obvious change points indicated by the dashed vertical lines. In each of segments, we take the average of the extrustion rate data as the extrusion rate in our simulation study, plotted in blue. In the regression, let $\boldsymbol{x_1} =$

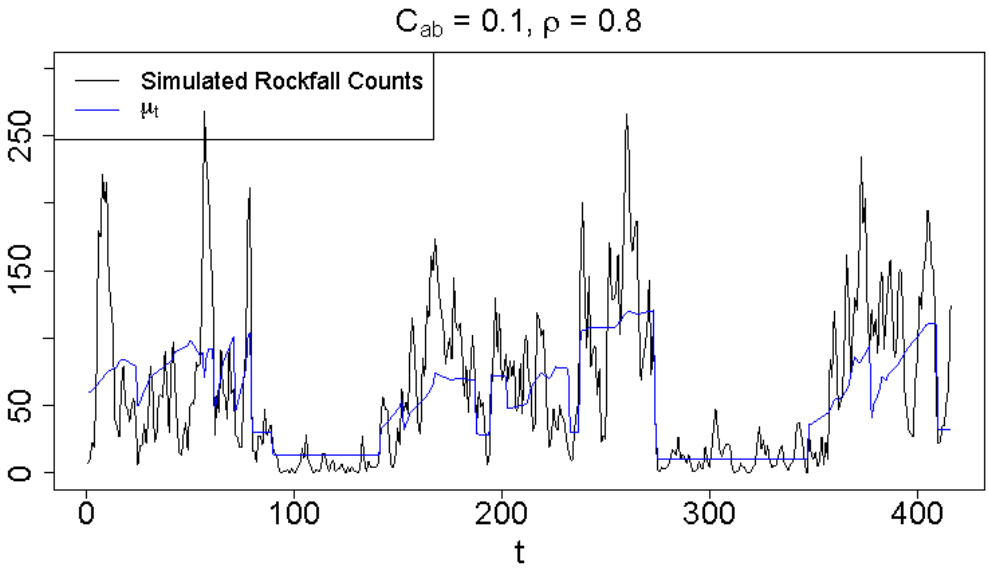


FIGURE 3.10: Simulation parameters $(\mu_t, C_{\alpha\beta}, \rho)$ and simulated rockfall counts y_t.

log(Extrusion rate + 1)and consequently, we can generate the mean process by

$$\log(\mu_t) = \gamma_0 + \gamma_1 \times x_{t1} + \gamma_2 \times x_{t2},$$

where we assume $\boldsymbol{\gamma} = (3, 0.7, 2.5)$. Then a dataset of 416 rockfall counts can be simulated as a dynamic negative binomial branching process, shown in Figure 3.10:

$$y_t \sim \text{bNB}(\mu_t, C_{\alpha\beta}, \rho), \quad t = 1, \ldots, T = 416,$$

where we assume $C_{\alpha\beta} = 0.1, \rho = 0.8$.

In the next section, we will show the simulation results. However, in the first place, we would like to test our new simulation method for the dynamic negative binomial branching process. Using the same set of simulation covariates/parameters $(\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{\gamma}, C_{\alpha\beta}, \rho)$, we generate 200 samples of $\{y_t\}_{t=1}^{T}$. We intend to examine the approximate marginal distributions, estimated sample lag-1 autocorrelation, as well as the estimated mean process. The simulation results are shown in Figures 3.11, 3.12 and 3.13.

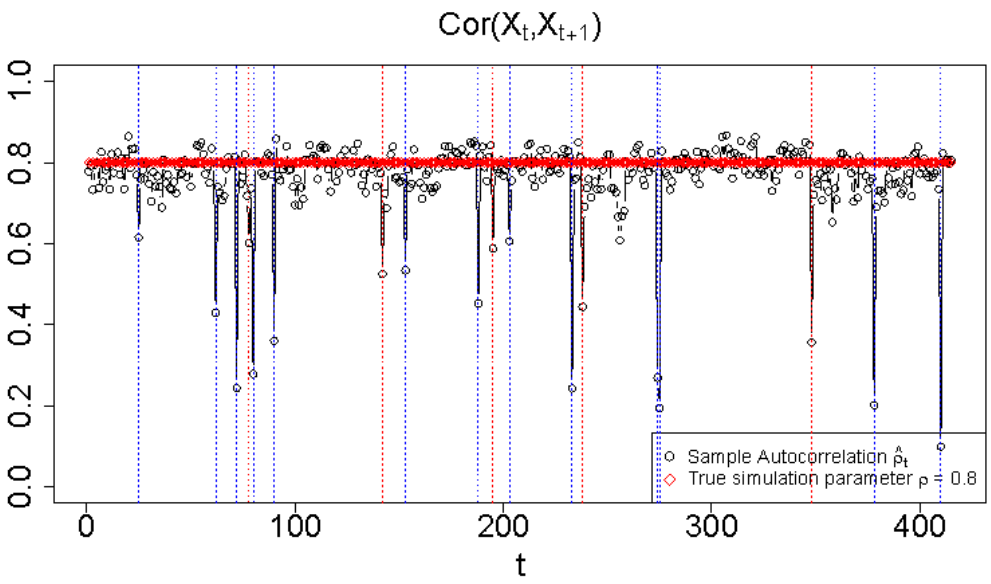


FIGURE 3.11: Simulation autocorrelation parameter ρ and estimated sample autocorrelation.

Black circles and red diamonds in Figure 3.11 represent lag-1 sample autocorrelation $\hat{\rho}_t$ and the true simulation parameter $\rho_t = \rho$ at each time point t, $t = 1, \ldots, T = 416$ respectively. It can be seen that most of the black circles gather around the true value at 0.8; however, at some points, the sample autocorrelation is much lower than 0.8, which can be explained by either abrupt increases (highlighted by red dashed vertical lines) or abrupt decreases (highlighted by blue dashed vertical lines) in the simulation parameter μ_t.

In each plot of Figure 3.12, the red curve is the "true" marginal distribution– negative binomial distribution with parameters listed under the histogram at a certain time point t. The histogram represents an approximation of the marginal negative binomial distribution with parameters estimated by the maximum likelihood method and denoted in the legend. Generally speaking, the histograms and the curves are good matches.

Figure 3.13 shows three kinds of processes. The red diamonds plot the mean

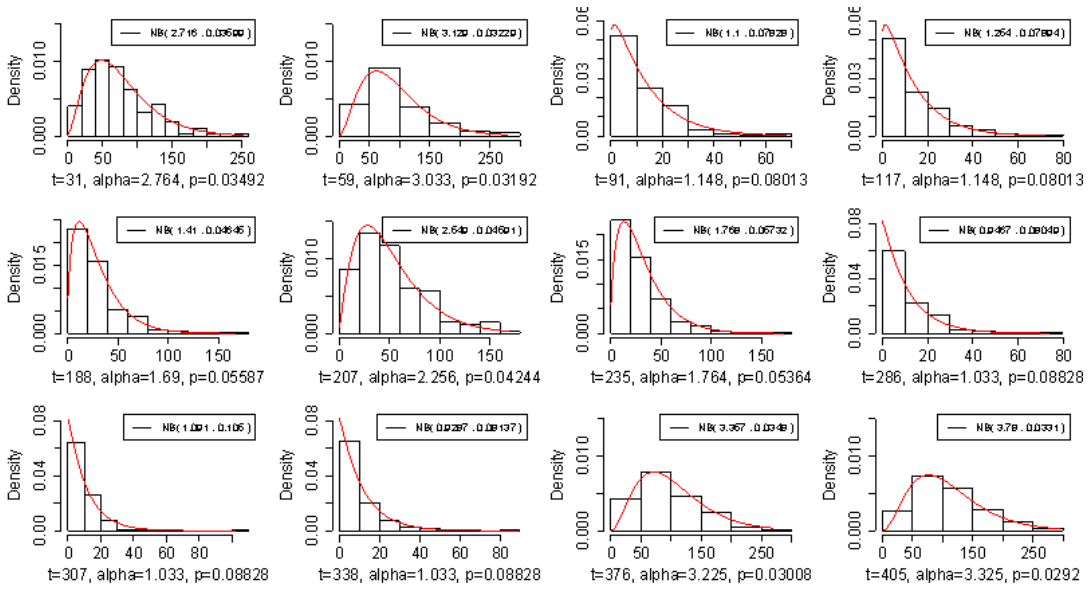


FIGURE 3.12: Approximate marginal distributions estimated from 200 samples at 12 random time points.

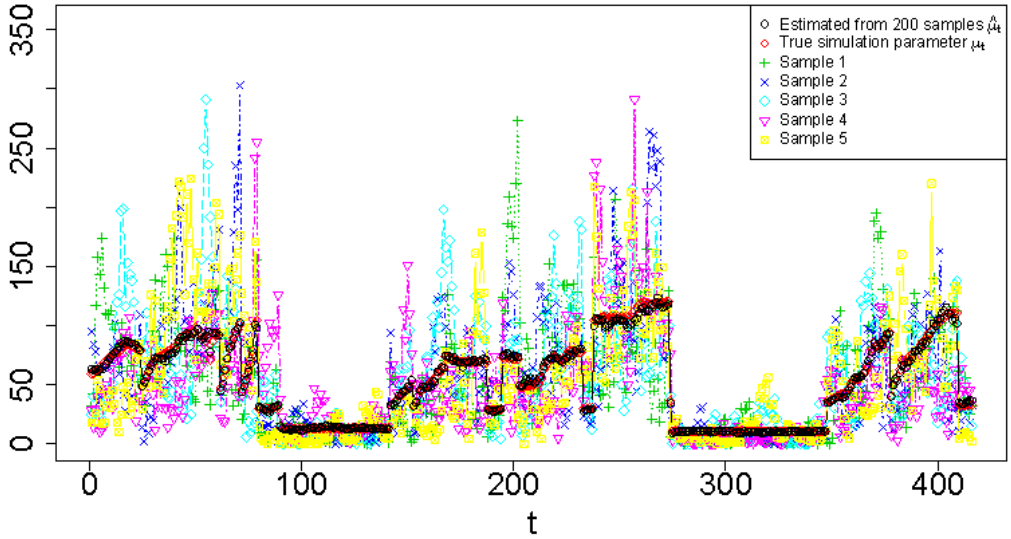


FIGURE 3.13: Simulation parameter–mean process μ_t.

process μ_t used to generate rockfall counts. The black circles are mean values estimated from the 200 sets of rockfall counts. The other colorful lines each represent one realization randomly selected from the 200 samples. It is straightforward to observe that the estimated mean process follows the true simulation parameter pretty closely, which together with the above two figures, indicate that our new simulation method for dynamic branching negative binomial processes works well.

3.5.5 Correlated Negative Binomial regression and posterior distributions

Now we proceed to the inference of the correlated negative binomial regression. To start with, let's take a look at the posterior distributions for the simulated dataset.

Figures 3.14–3.18 present the posterior analysis of the simulation parameters $\boldsymbol{\gamma}$, $C_{\alpha\beta}$, ρ and μ_t. In the first three figures, the traceplots, autocorrelation functions and histograms (all in black) show trajectories, autocorrelations of a series of time lags of the original 100,000 MCMC samples, and approximate marginal densities of 400 MCMC samples after burnin (20,000) and thinning (200); the red dashed lines represent the true values of the parameters in the simulation; and the red dotted lines plot the prior distributions of one of the three parameters $\boldsymbol{\gamma}$, $C_{\alpha\beta}$ or ρ.

All the posterior samples look fine, in the sense that the MCMC samples are close to the true simulation parameter values and the autocorrelation of the decays to 0 within lag 200. At first sight from Figure 3.16, it seems that the posterior samples of ρ differ significantly from the true simulation value 0.8; however, according to the way we deal with λ_t when abrupt drops occur in our simulation method, together with Figure 3.11, it actually makes prefect sense that the lag-1 autocorrelation is lower than the true value.

Figure 3.17 illustrates the posterior mean (black solid line) and the 90% credible intervals (blue dashed lines) for the mean process μ_t and compares it with the true simulation parameter plotted in green solid line and simulated rockfall counts in red.

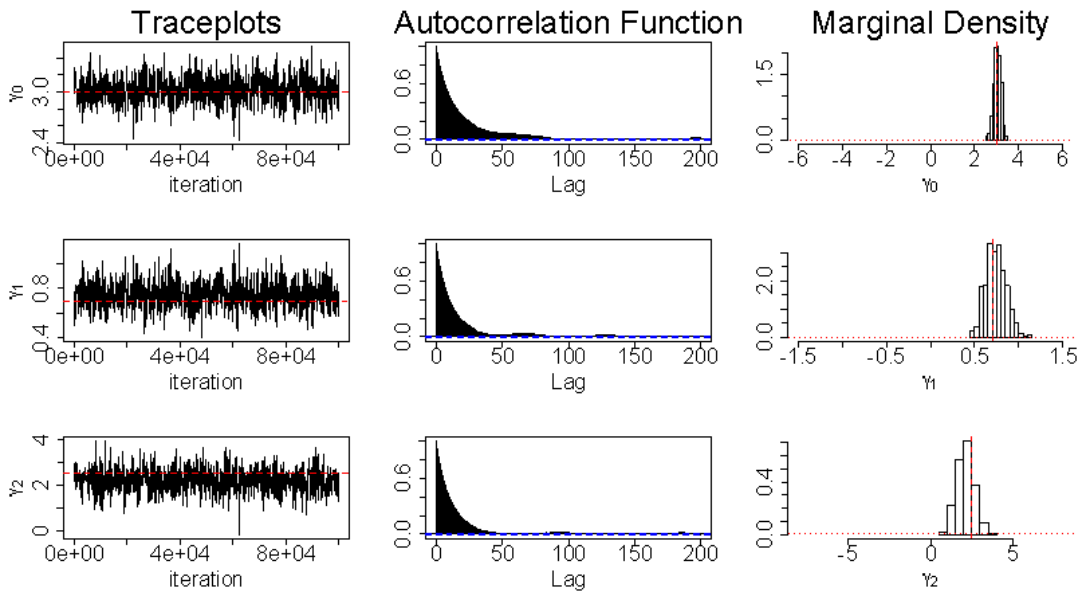


FIGURE 3.14: Simulation Study: Traceplots, autocorrelation function, and posterior marginal distributions for $\boldsymbol{\gamma} = \{\gamma_0, \gamma_1, \gamma_2\}$.

The real coverage indicates the proportion of simulated data covered by the credible intervals and the result is satisfactory.

Figure 3.18 shows the correlation between posterior samples of the parameters. There is an obvious linear correlation between γ_0 and γ_1, because the regression covariate x_1 was not centralized.

Figures 3.19–3.23 present similar results to those above for the real data on rockfall counts, dome height and extrusion rate, except that no true values of these parameters can be plotted in the figures to compare with the posterior samples.

3.6 Inverse problem: inference on the extrusion rate

Tracking extrusion rate is central both to science and to prediction of volcanic hazards, yet it cannot be measured directly. The ultimate goal of this study was to find a method to estimate extrusion rate from the easily measured variables of dome height and rockfall counts; this is an inverse problem.

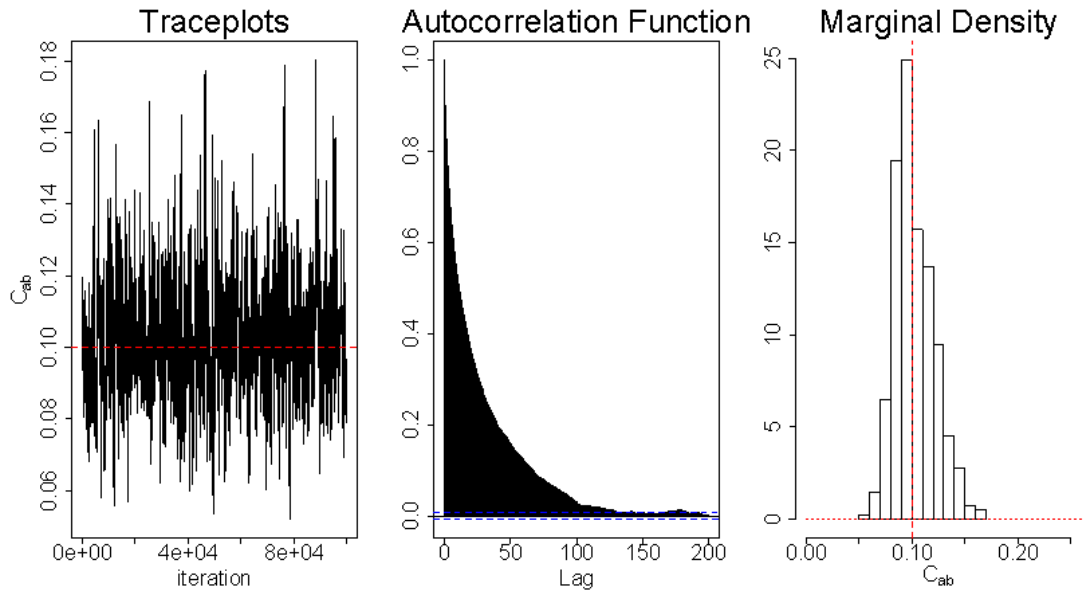


FIGURE 3.15: Simulation Study: Traceplots, autocorrelation function, and posterior marginal distributions for $C_{\alpha\beta}$.

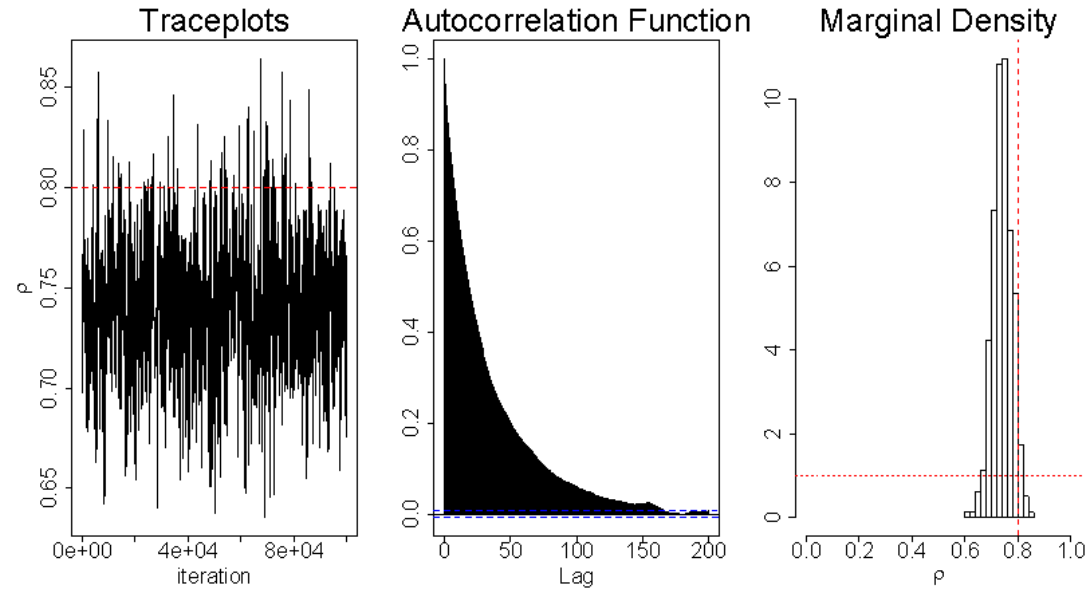


FIGURE 3.16: Simulation Study: Traceplots, autocorrelation function, and posterior marginal distributions for ρ.

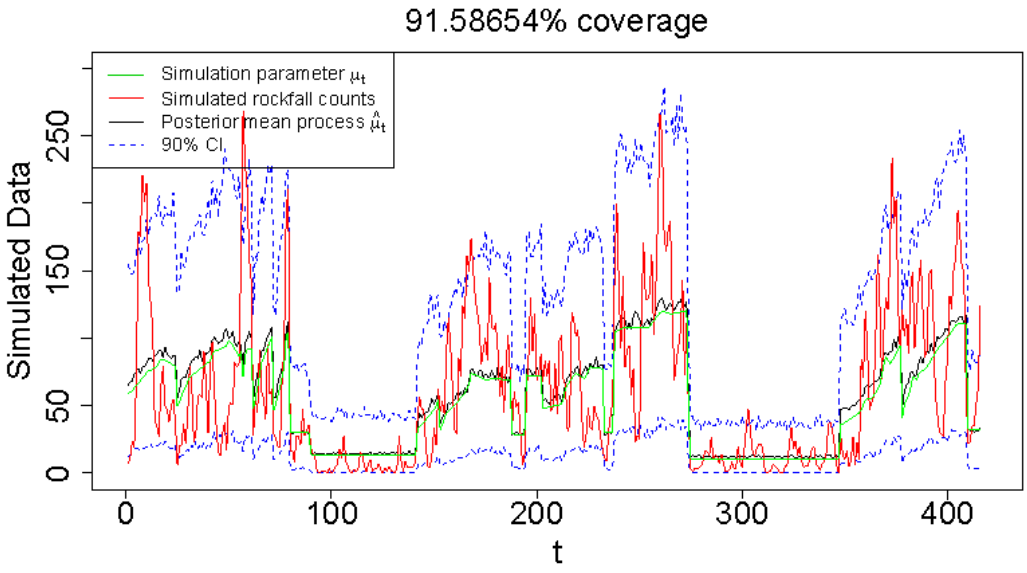


FIGURE 3.17: Simulation Study: Posterior mean and credible intervals compared with true simulation parameter and data.

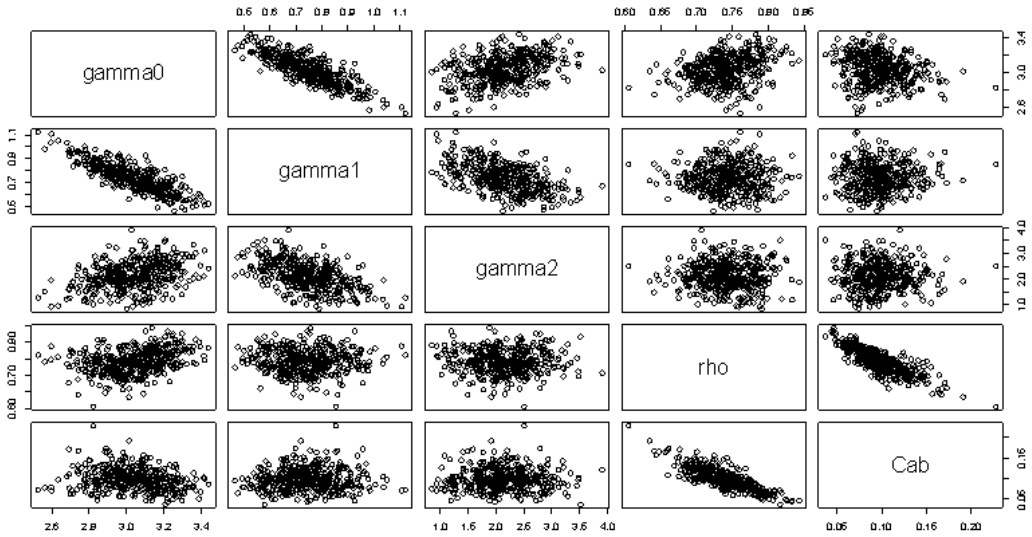


FIGURE 3.18: Simulation Study: pairwise correlation between posterior samples.

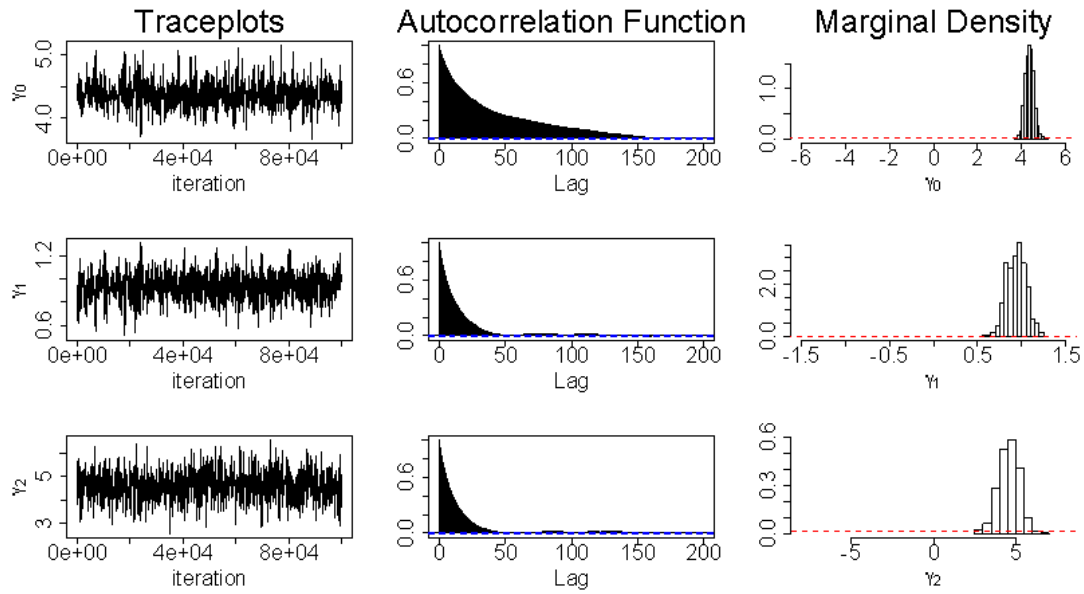


FIGURE 3.19: Real data: Traceplot, autocorrelation and posterior distribution for $\boldsymbol{\gamma}$.

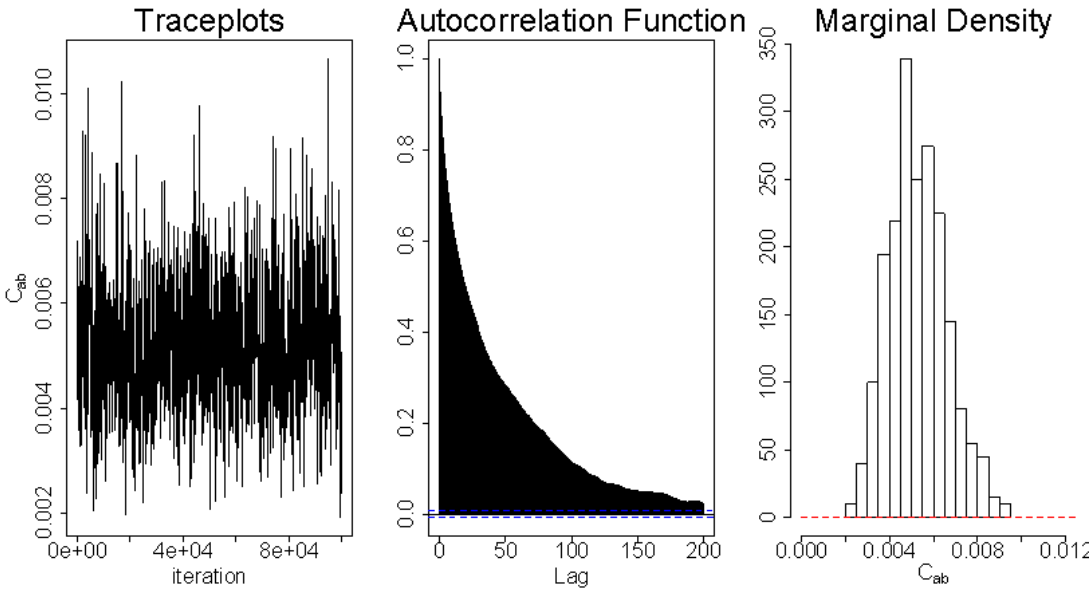


FIGURE 3.20: Real data: Traceplot, autocorrelation and posterior distribution for $C_{\alpha\beta} = \alpha_t \beta_t$.

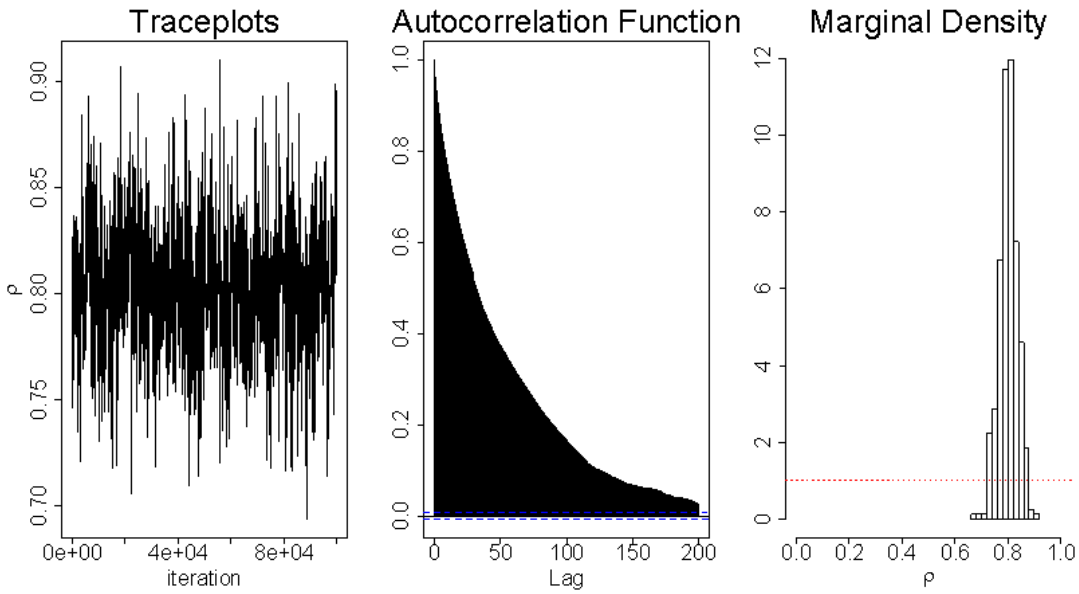


FIGURE 3.21: Real data: Traceplot, autocorrelation and posterior distribution for ρ.

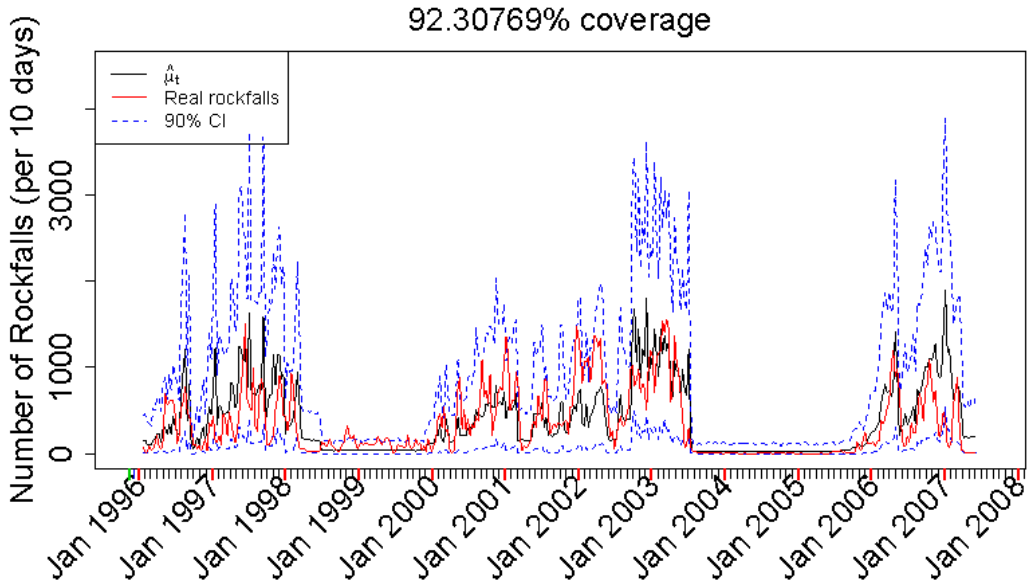


FIGURE 3.22: Real data: Posterior mean of the process and its 90% credible interval. Red line plots the real rockfall counts. Black line is the estimated mean process. Dashed lines represent 90% credible intervals; they in fact cover 92.30769% of the real data.

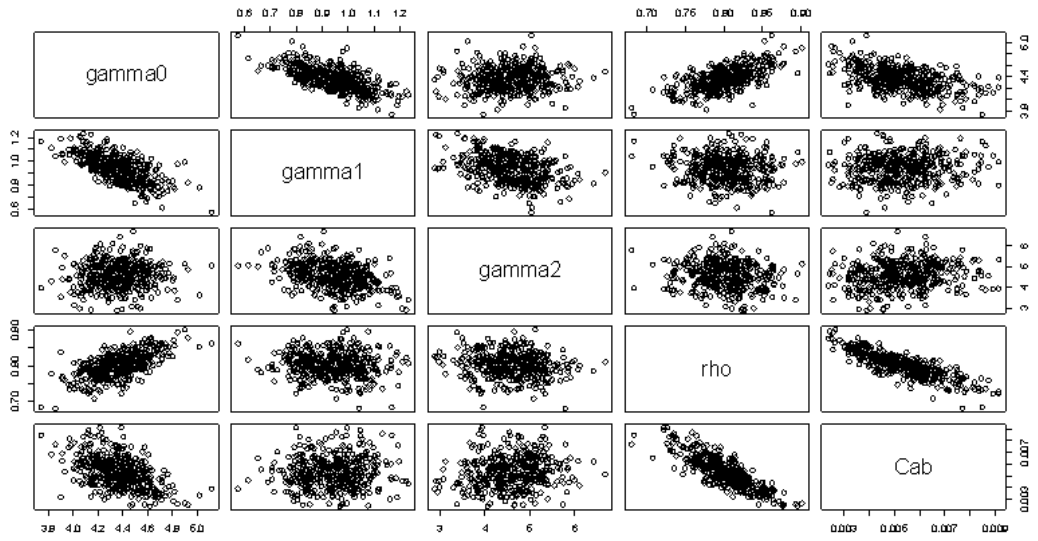


FIGURE 3.23: Real data: Pairwise correlation of posterior samples.

As in the previous regression model, we denote the extrusion rate by $\boldsymbol{x_1}$. Under the assumption of the correlated negative binomial regression model, the extrusion rate of day t can be viewed as a function of the dome height and number of rockfalls of the same day, together with the posterior estimates of model parameters: $\boldsymbol{\gamma}$, $C_{\alpha\beta}$ and ρ.

The inverse problem is formulated as follows:

$$x_{t1} = f(x_{t2}, y_t, \widehat{\gamma_0}, \widehat{\gamma_1}, \widehat{\gamma_2}, \widehat{C_{\alpha\beta}}, \hat{\rho}), \qquad \text{where } f \text{ is an unknown function}, t = 1, \ldots, T.$$

1. The likelihood function is the same function as used for the negative binomial regression model, while for the inverse problem, only the extrusion rate process will be considered as a "parameter", and all other information as "data".

$$L(\mathbf{x_1} \mid \mathbf{y}, \mathbf{x_2}, \widehat{\gamma_0}, \widehat{\gamma_1}, \widehat{\gamma_2}, \widehat{C_{\alpha\beta}}, \widehat{\rho})$$

$$= \prod_{t=1}^{T} \text{bNB}(\boldsymbol{y_t} \mid \exp(\widehat{\gamma_o} + \widehat{\gamma_1} \times x_{t1} + \widehat{\gamma_2} \times x_{t2}), \widehat{C_{\alpha\beta}}, \widehat{\rho}),$$

where $\widehat{\boldsymbol{\gamma}} = (\widehat{\gamma}_0, \widehat{\gamma}_1, \widehat{\gamma}_2)$, $\widehat{C_{\alpha\beta}}$ and $\widehat{\rho}$ are posterior modes from the correlated negative binomial regression in the previous section.

2. Prior distribution for f.

   We assume f is a step function. Figure 3.24 shows $k+1$ piecewise constant values $\{a_i\}$ over disjoint intervals $\{s_{i-1}, s_i\}$, where $i = 1, \ldots, k+1, s_0 = 0, s_{k+1} = T$. Similar to what we did in Chapter 2, we assume that the number of change points follows a Poisson distribution with mean $T\delta$ and add a second level of hierarchy to the parameter δ; the positions of the k change points are the order statistics uniformly distributed on $[0, T]$. We assume uninformative proper prior distribution for $\{a_i\}_{i=1}^k$ rather than improper hyperprior distributions, because the posterior distribution is too complicated for propriety checking. In our model, the $\{a_i\}_{i=1}^k$ are are assumed to be positive, although in reality they can be zero, so we assign a Gamma density (see Figure 3.25) which has certain mass around zero while being flat at other values.

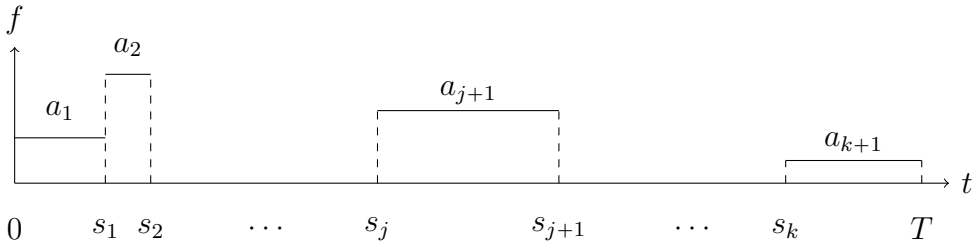


FIGURE 3.24: Illustration of a step function f.

The priors are listed as below:

$$\delta \sim \pi(\delta) \propto 1/\sqrt{\delta} 1_{[0,1]},$$

$$(k \mid \delta) \sim \text{Poi}(T\delta),$$

$$(s_1, \ldots, s_k | k) \sim \text{Unif}(0, T),$$

$$(a_1, a_2, \ldots, a_{k+1} | k, \alpha_a, \beta_a) \overset{\text{iid}}{\sim} \text{Gamma } (\alpha_a = 0.5, \beta_a = 0.2).$$

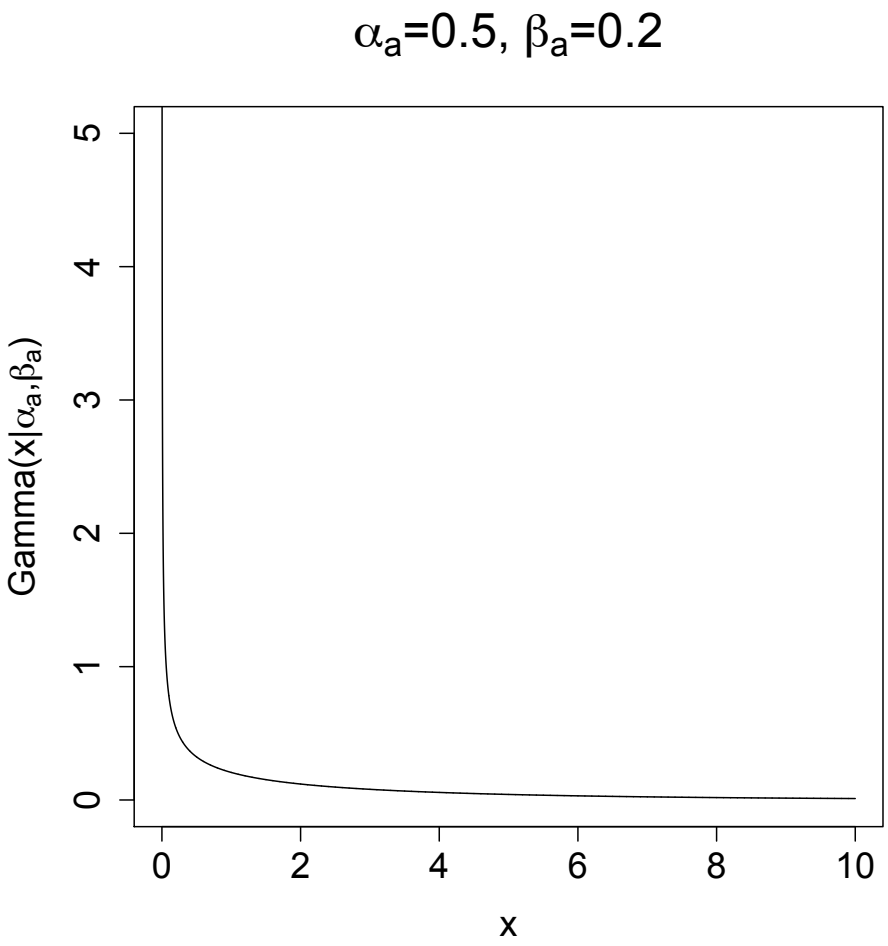


FIGURE 3.25: Gamma prior for $\{a_i\}_{i=1}^k$.

3.6.1 Simulation results

Firstly, we apply the above method to our simulated dataset. Figures 3.26–3.28 respectively illustrate the posterior distribution of the number of change points, k, posterior samples of positions of change points, $\boldsymbol{s}$, and posterior mean and credible intervals for the rates. As shown in Figure 3.9 or the top plot in Figure 3.27, while we assumed there were 8 change points in the time series of extrusion rate in the simulation, it seems from the posterior results that the sampler found the big changes but missed the 3rd-6th change points, which represent two changes occurring on short

intervals. However, the estimated extrusion rate is pretty close to the one used in the simulation, as can be seen in Figure 3.28.

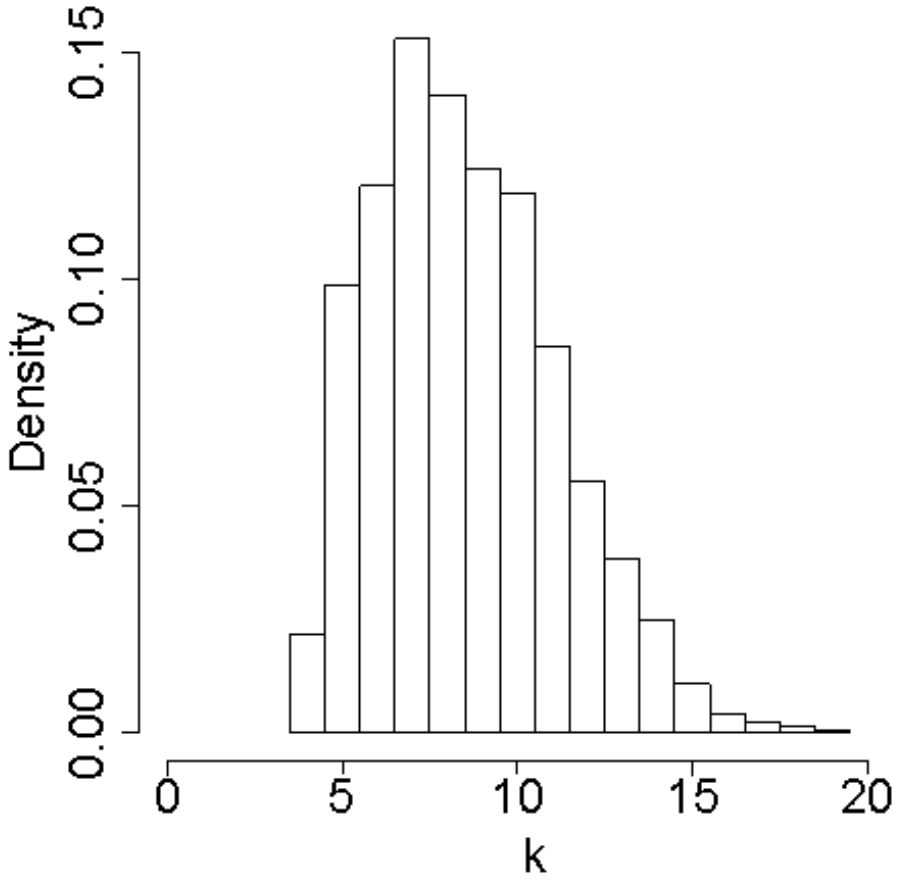


FIGURE 3.26: Simulation Study: Posterior distribution of k, the number of change points.

3.6.2 Real data

When the method was applied to the real data, we obtained posterior results having similar features to the simulation study. For the extrusion rate, the sampler identified only a few significant change points, whereas the extrusion rate data had many more change points. (But, again, that data was itself only inferred data, and not reality.)

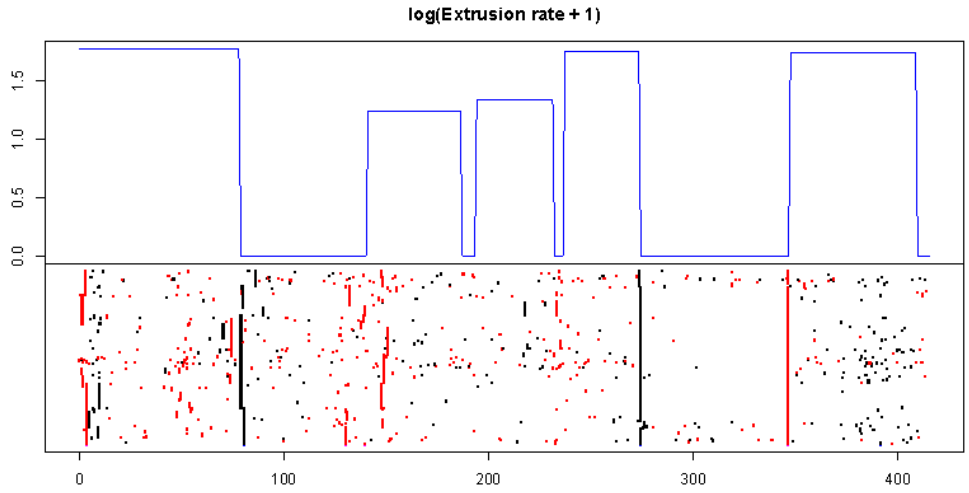


FIGURE 3.27: Simulation Study: Posterior samples of $\boldsymbol{s}$, positions of change points.

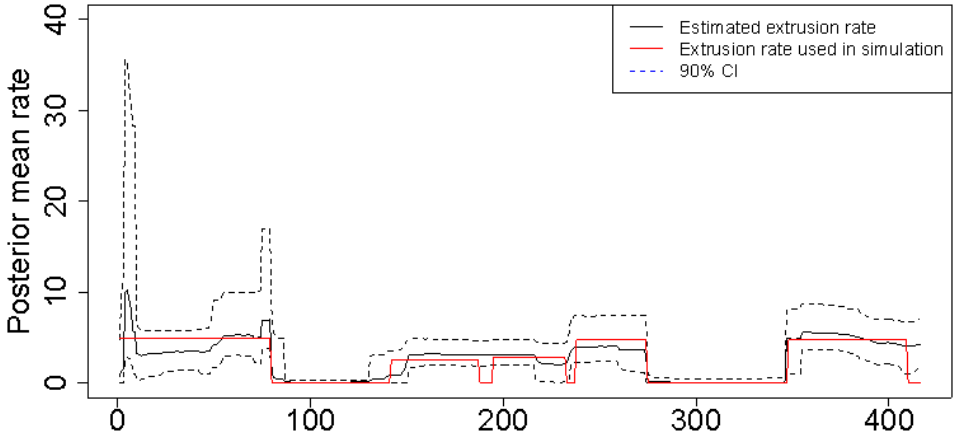


FIGURE 3.28: Simulation Study: Posterior estimate and credible intervals of the step function of extrusion rate, compared with the true extrusion rate used in simulation.

As shown in Figure 3.29, the posterior mode of the number of change points is 6. In Figure 3.30, the posterior samples of $\boldsymbol{s}$ form vertical lines at times when there is an obvious change in the extrusion rate and the change lasts for a relatively long period of time. The estimated extrusion rate presented in Figure 3.31 does not match the real data as well as in the simulation study, especially around January 2007. However, this can be explained by the discrepancy between $\hat{\mu}$ and the true parameter value around January 2007 in Figure 3.22, which implies a lack of fit in that time period, and thus bad estimates plugged into the model for the inverse problem. (In this time period, there were high extrusion rates and dome heights, yet low rockfalls – see Figure 3.8 – something that implies there is some missing science in the model.) Nevertheless, the lack of fit in the regression does not indicate that our correlated Negative Binomial regression model itself is not working well.

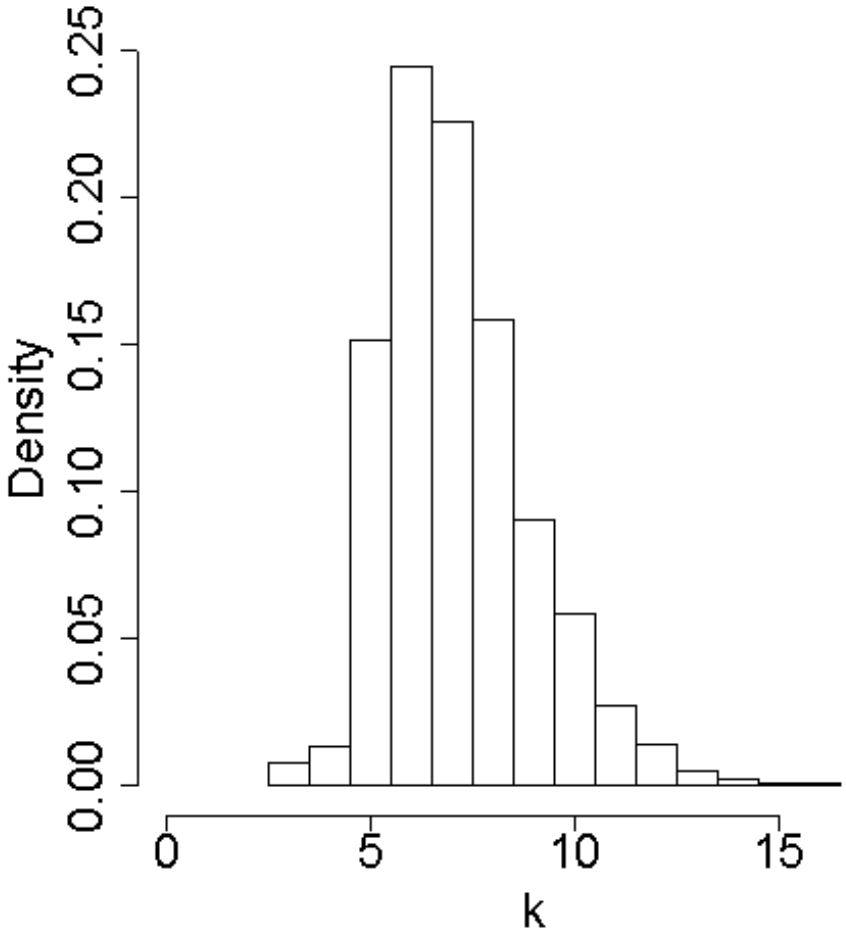


FIGURE 3.29: Real data analysis: Posterior distribution of k, the number of change points.

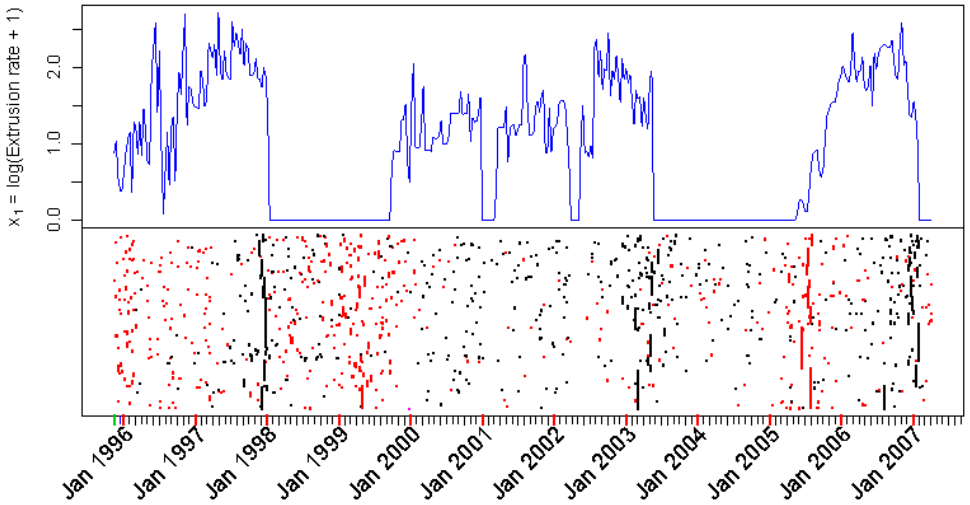


FIGURE 3.30: Real data analysis: Posterior samples of $\boldsymbol{s}$, positions of change points.

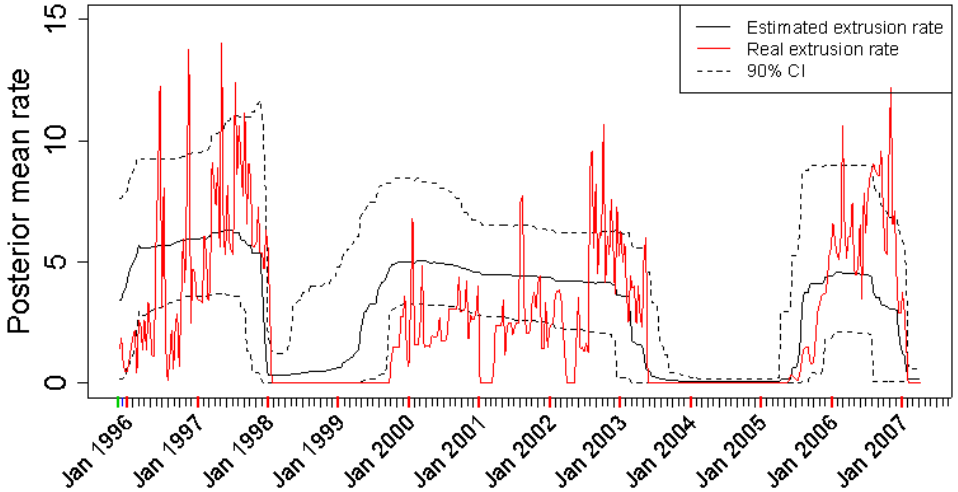


FIGURE 3.31: Real data analysis: Posterior estimate and credible intervals of the extrusion rate, compared with the real extrusion rate.

4

Adaptive Energy Partitioning for Generalized Wang-Landau Sampling

4.1 Introduction

Generating samples from a complex distribution which is known up to a normalizing constant is a ubiquitous and often challenging problem in Bayesian computation. Markov Chain Monte Carlo (MCMC) methods are commonly used for this purpose. But they often fail when the target distribution is multimodal, getting trapped in local modes.

The Wang-Landau (WL) algorithm (Wang and Landau, 2001) is an adaptive sampling scheme that modifies the target distribution to enable the chain to visit low-density regions of the state space. While the original algorithm was developed for discrete state space arising in statistical physics, its impressive performance has led to extensions for general state space arising in Bayesian statistical inference, referred to as generalized Wang-Landau (GWL) algorithm in this chapter. For example, Liang (2005) applies a stochastic approximation approach, and Atchade and Liu (2010) provide convergence results under regularity conditions. Both studies focus

on the choice of a step size sequence discussed later in Section 4.2, but rely heavily on user-specified partition of the state space. This makes implementation and use of the algorithm more time-consuming and less automatic, and will drastically affects the performance of the algorithm.

In this chapter, we develop an automatic, adaptive partitioning scheme which continually refines the initial partition as needed during sampling. By doing so, we overcome the limitations of the user-specified input partition, making the algorithm significantly more automatic and user-friendly. The performance of the algorithm also becomes more reliable and robust for exactly those multimodal problems which WL/GWL sampling is designed to address.

4.2 The Wang-Landau and generalized Wang-Landau algorithms

We first briefly review the Wang-Landau (WL) algorithm. Let $\pi(x)$ be a distribution defined over a finite state space $\mathcal{X}$ and known up to a constant c,

$$\pi(x) = h(x)/c, \quad x \in \mathcal{X}.$$

Let $E(x) = -\log(h(x))$ denote the "energy" function and $\{E_1, \ldots, E_d\}$ be the set of real numbers containing all possible values of $E(x)$. Define ϕ_i be the number of states that are mapped to E_i under E,

$$\phi_i = \phi(E_i) = |\{x : E(x) = E_i\}|$$

and $\phi = (\phi_1, \ldots, \phi_d)$ be the density of the states. The goal of the WL algorithm is to estimate ϕ, achieved by incrementally estimating ϕ and at the same time using the estimates to modify the target distribution to be uniform over the allowed energy range (a "flat histogram"), enabling the crossing of energy barriers.

Let $\hat{\phi}_i$ denote the working estimate of ϕ_i. A run of the WL algorithm begins with initial estimates

$$\hat{\phi}_1 = \cdots = \hat{\phi}_d = 1.$$

In each subsequent iteration, a sample x^* is simulated by a single Metropolis update with invariant distribution

$$\pi\hat{(x)} \propto 1/\phi(\hat{E}(x))$$

and set

$$\hat{\phi}_i = \hat{\phi}_i \delta^{\mathbf{1}(E(x^*)=E_i)}, \quad 1 \leqslant i \leqslant d,$$

where $\delta > 1$ is a modification factor and $\mathbf{1}(\cdot)$ is the indicator function. The algorithm iterates until a flat histogram has been produced in the energy space. Once the histogram is flat, the algorithm reduces the initial values $\hat{\phi}$ and δ according to a predefined scheme and restarts. The simulation stops when δ is very close to 1, say

$$\log(\delta) \leqslant 10^{-8}.$$

Notice that many sampling problems of interest in statistics and statistical mechanics involve continuous state spaces. In the following, we describe an extension of the Wang-Landau algorithm to the generate state spaces (Atchade and Liu, 2010).

Let $(\mathcal{X}, \mathcal{B}, \lambda)$ be a countably generated measure space, where λ is a σ-finite measure. Let $\pi(dx) \propto \pi(x)\lambda(dx)$ be a probability measure on $\mathcal{X}$, $(\mathcal{X}_i)_{1\leqslant i\leqslant d}$ be a partition of $\mathcal{X}$ along the energy function $-\log \pi(x)$, and $\phi(i) = \pi(\mathcal{X}_i)$.

The generalized Wang-Landau (GWL) algorithm attempts to simultaneously estimate the marginal probabilities $\{\phi(i)\}$ and sample from the modified target distribution obtained by reweighing each element inversely proportional to its marginal probability, in order to spend equal time in each component (see Algorithm 4.2.1).

Algorithm 4.2.1 (Generalized Wang-Landau)**.** *Let $\{\gamma_n\}$ be a decreasing sequence of positive numbers (e.g. $\{1/n\}$). Given $X_0 \in \mathcal{X}$, find the partition element I_0 to which X_0 belongs. Set $a_0 = 0$, $\kappa = 0$, $\epsilon \in (0, 1)$, and $\phi_0 \in \mathbb{R}^d$ such that $\phi_0(i) > 0$, $i = 1, \ldots, d$. At each iteration $n \geqslant 0$, given $X_n \in \mathcal{X}$, $I_n \in \{1, \ldots, d\}$, $\phi_n \in \mathbb{R}^d$, a_n and κ:*

1. *Sample X_{n+1} from a transition kernel P_ϕ with invariant distribution*

$$\pi_{\phi_n} \propto \sum_{i=1}^{d} \frac{\pi(x)}{\phi_n(i)} \mathbf{1}_{\mathcal{X}_i}(x)$$

2. *For $i = 1, \ldots, d$, set $\phi_{n+1}(i) = \phi_n(i)\left(1 + \gamma_{a_n} \mathbf{1}_{\{I_{n+1}=i\}}\right)$.*

3. *If $\max_i |v_{\kappa,n+1}(i) - 1/d| \leqslant \epsilon/d$, then set $\kappa = n+1$ and $a_{n+1} = a_n + 1$, otherwise set $a_{n+1} = a_n$, where $v_{k,n}(i) = \dfrac{1}{n-\kappa} \displaystyle\sum_{j=k+1}^{n} \mathbf{1}_{\{I_j=i\}}$.*

Atchade and Liu (2010) points out that the performance of the GWL algorithm depends on the choice of the partition into subspaces. Kou et al. (2006) recommends a reasonable heuristic by fixing the lowest and second highest energy levels and assigning the other energy levels by a geometric progression (equivalent to making $\log(E_i)$ evenly spaced). That is, partition $\mathcal{X}$ into d component $\{\mathcal{X}_i\}_{i=1}^d$

$$\mathcal{X}_i = \{x \in \mathcal{X} : E_{i-1} \leqslant E(x) \leqslant E_i\},$$

where

$$E_0 = E_{\min}, \quad E_1 = E_0 r_e, \quad \ldots, \quad E_{d-1} = E_0 r_e^{d-1} = E_{\max}, \quad E_d = \infty,$$

for some geometric rate r_e, typically computed after choosing $E_{\min}$, $E_{\max}$ and d.

It is worth mentioning that the convergence behavior of WL and GWL algorithms can strongly depend on the choice of the partition. We illustrate this using a simple bimodal target distribution, a mixture of two normal distributions in two dimensions (Figure 4.2a)

$$\pi(\mathbf{x}) = \frac{1}{2}\left[N(\mathbf{x}; (-5,-5), I) + N(\mathbf{x}; (5,5), I)\right] = h(\mathbf{x})/2. \tag{4.1}$$

As $h(\mathbf{x}) \leqslant 1 + e^{-100}$, the energy function $E(\mathbf{x})$ satisfy

$$E(\mathbf{x}) = -\log(h(\mathbf{x})) \geqslant -\log(1 + e^{-100})$$

and it can be used to set $E_{\min}$ of the initial energy partition. Moreover, fix

$$E_{\max} = -\log(10^{-10})$$

by a rough estimate of the lowest density value. For this problem, a user might reasonably suspect that few energy levels are need, say $d = 4$. Set $\gamma_n = 1/n$ and $\epsilon = 0.3$ as recommended by Atchade and Liu (2010). The random walk proposal has step size 1 and the simulation is started from one of the two modes, $(5, 5)$.

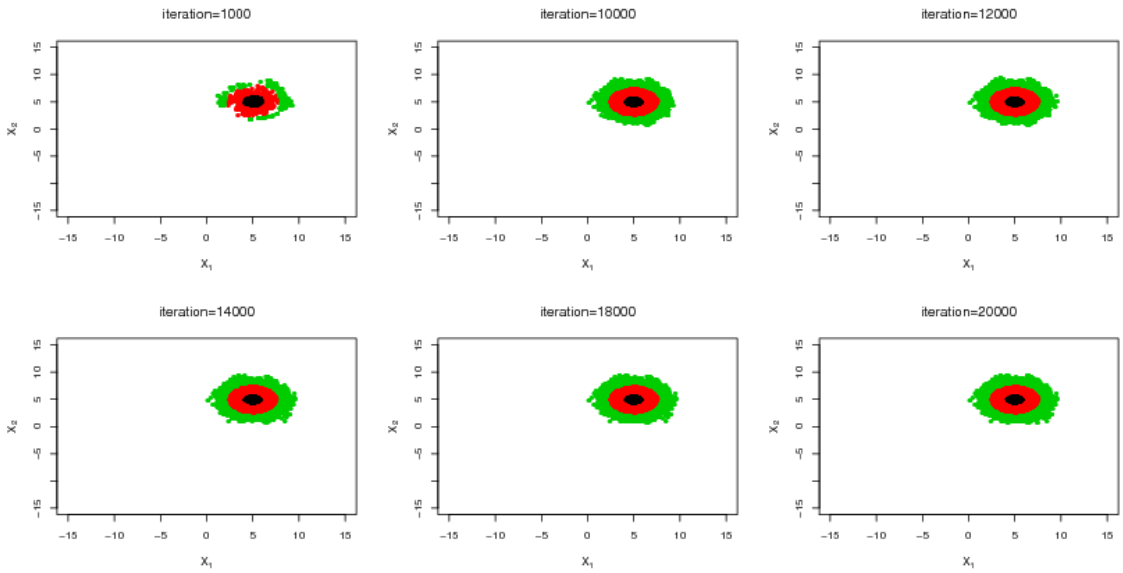


FIGURE 4.1: Sample path of GWL algorithm on two-component normal mixture distribution (4.1) with modes at $(-5, -5)$ and $(5, 5)$ for $20,000$ iterations. The chain never escapes the mode in which it was initialized, failing to cross the energy barrier to the other mode.

Figure 4.1 shows the resulting sample history, with observed samples falling into each distinct partition element assigned a different color. We see that the chain visits only three of the four subspaces, and remains trapped in the mode in which it was initialized.

To see why this occurs, Figure 4.2c shows the energy partition along with the known unnormalized density $h(x)$ for a slice through the state space along the axis

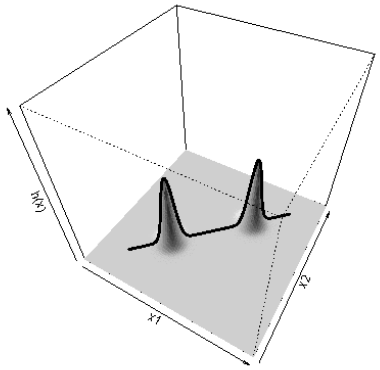
between modes (shown in black in Figure 4.2a). Colors below the x-axis in Figure 4.2c indicate the locations of the partition elements corresponding to the rings in Figure 4.1. Figure 4.2b shows the density along this axis, within the third energy bin. The GWL sampler aims to sample energy bins uniformly, but the weights applied to achieve this are constant *within* a given energy bin. As a result, the sampler visits points within the bin according to the unmodified target distribution $\pi(x)$ restricted to the energy interval. From Figure 4.2b we immediately see the problem: the density decays exponentially in E, so if $|E_{i+1} - E_i|$ is large *relative to the scale of the Metropolis proposal kernel*, the chain will visit the ith bin but spend exponentially small time in the region where a move to the $(i+1)$-st bin can be proposed. As a result, even when the $\phi(i)$'s are estimated perfectly, the conductance of the chain will be exponentially low in the ratio $\sigma/|L_i|$, where σ is the proposal variance and L_i is the length of the set $\mathcal{X}_i$ along the axis between modes.

The failure of the GWL algorithm to cross over the energy barrier to other modes is particularly troublesome since that is precisely the problem the algorithm is designed to address.

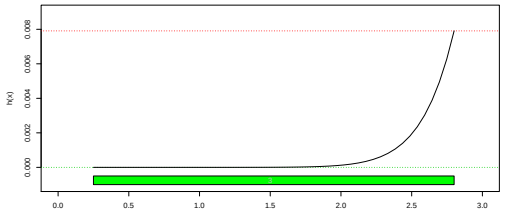
4.3 Adaptive energy partitioning

The strong dependence of convergence behavior of WL and GWL algorithms on the choice of energy partition parameters limits the general utility of these methods. Here we describe a fully automatic algorithm for adaptively updating the energy levels during the course of the algorithm, which allows the algorithm to overcome the limitation of the initially-specified partition. The resulting behavior is therefore user-independent. In the rest of the chapter, we refer to this algorithm as Adaptive Energy Generalized Wang-Landau (AE-GWL) algorithm.

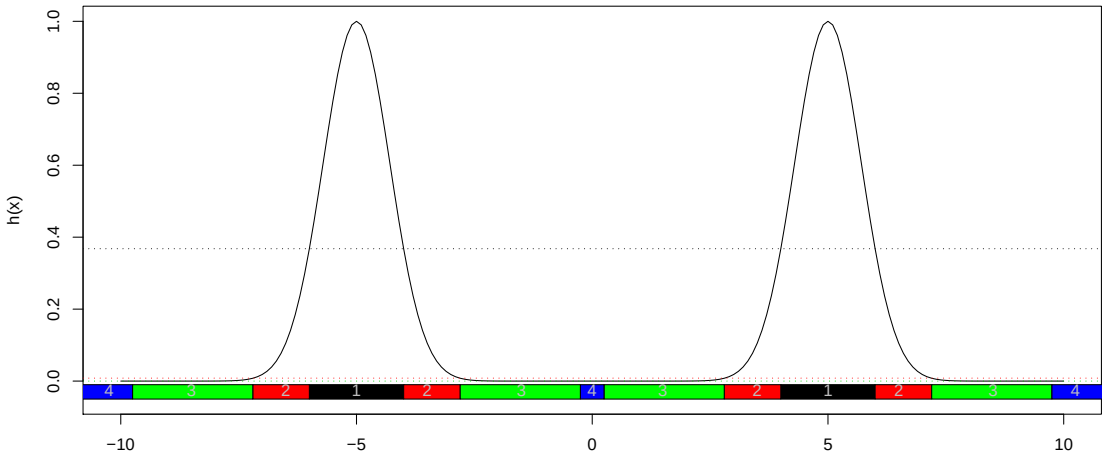
Algorithm 4.3.1. *Add to Algorithm 4.2.1:*



(a) Density for normal mixture (4.1).



(b) Closer view of the target density over the region $\mathcal{X}_3 \subset \mathcal{X}$ specified by energy bin 3. The density concentrates strongly on the right-hand side.



(c) Energy partition obtained by geometric progression, shown along axis between modes. Colored regions show locations of two-dimensional rings given by energy partition, with length indicating the corresponding ring width.

FIGURE 4.2: Density and energy rings for the bimodal distribution example.

1. *Initialization by a geometric progression:*

$$E_0 = \inf_x E(x), \quad E_1 = \max\{E_0 r_e, 1\}, \quad E_2 = E_1 r_e,$$

$$\ldots, \quad E_{d-1} = E_1 r_e^{d-2}, \quad E_d = \infty$$

2. *Every* n_{split} *iterations, if*

$$|\log(\hat{\phi}_i) - \log(\hat{\phi}_{i+1})| \geqslant E_{th}$$

for energy threshold E_{th} *and some* i*, subdivide the* i*th energy ring as follows:*

$$E_0, E_1, \ldots, E_i, E_{i+1}^* = \sqrt{E_{i+1}E_i}, E_{i+1}, \ldots, E_{d-1}, E_d.$$

(When $E_0 = 0$ *and* $E_1 = 1$*, the first energy ring can also be split in a similar way by adding* $E_1^* = E_1^2/E_2$*, which makes* $E_1^*/E_1 = E_1/E_2$*.)*

Note that we again use geometric progression in splitting the energy bins.

We re-apply the GWL algorithm in combination with the adaptive energy level refinement to the bimodal problem. The GWL algorithm uses the same initial parameters as those in Figure 4.1. Parameters E_{th} and n_{split} are chosen to be 5 and nIter/100, respectively. The new results are shown in Figure 4.3. We see that the chain easily crosses the energy barrier and discovers the other mode in less than $10,000$ iterations. At the end of $20,000$ iterations, the energy levels have been refined from the initial value of four to 13.

Notice that the rough estimate of $E_{\max} = -\log(10^{-10})$ turns out to work well for the bimodal example. In general, it is typically difficult to obtain a good estimate for a more complex problem. If the barrier is higher than expected, updating only the internal energy levels can fail to address this, which is another potential limitation of the user-specified partition. To see this, we instead choose $E_{\max} = -\log(10^{-3})$ for the bimodal example. Figure 4.4 presents the sample history of the chain in this

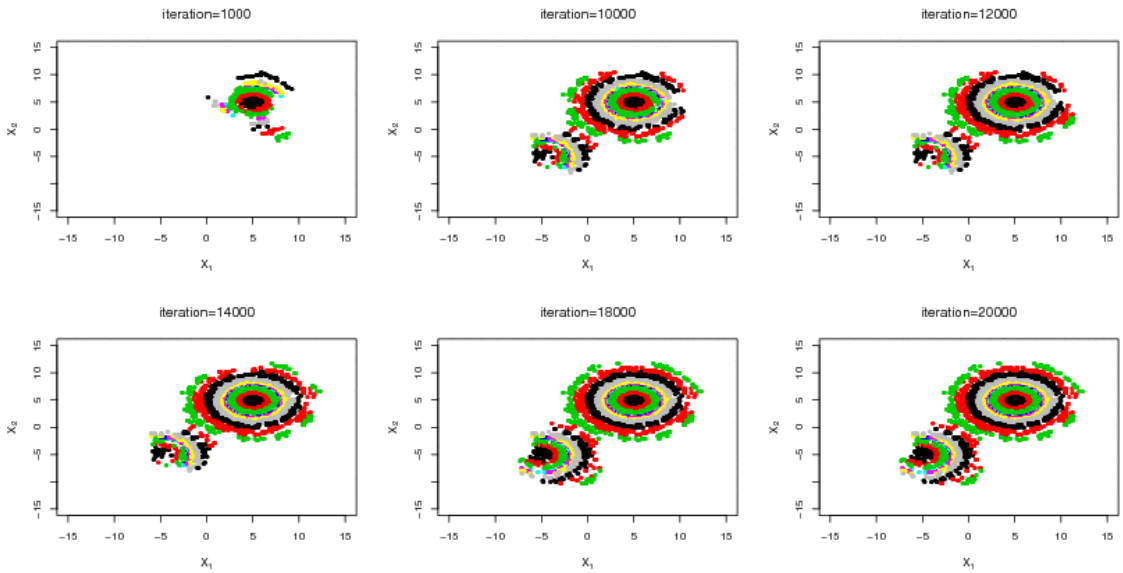


FIGURE 4.3: Sample path of adaptive energy partitioning GWL algorithm applied to the bimodal target distribution. Automatic refinement of the energy partition enables the chain to escape from energy bin 3 and successfully cross the barrier to the other mode in less than $10,000$ iterations.

case. Although the initial $d = 4$ energy rings have been split into five rings, indicating that the internal energy levels are split when necessary, the chain nevertheless fails to escape the mode in which it is initialized. Figure 4.5 shows why this occurs.

It is now clear that the performance of the WL algorithm also depends on the choice of $E_{\max}$. In order to make the algorithm more robust to the initial specification of $E_{\max}$, we modify the algorithm to take an arbitrary initial value and update it adaptively.

Algorithm 4.3.2. *Add to Algorithm 4.3.1:*

3. Every n_{split} iterations, also update the second highest energy level:

$$E_0, E_1, E_2, \ldots, E_{d-1}, E^*_{d-1} = \frac{E^2_{d-1}}{E_{d-2}}, E_d = \infty.$$

Here we again use geometric progression to split the energy bin.

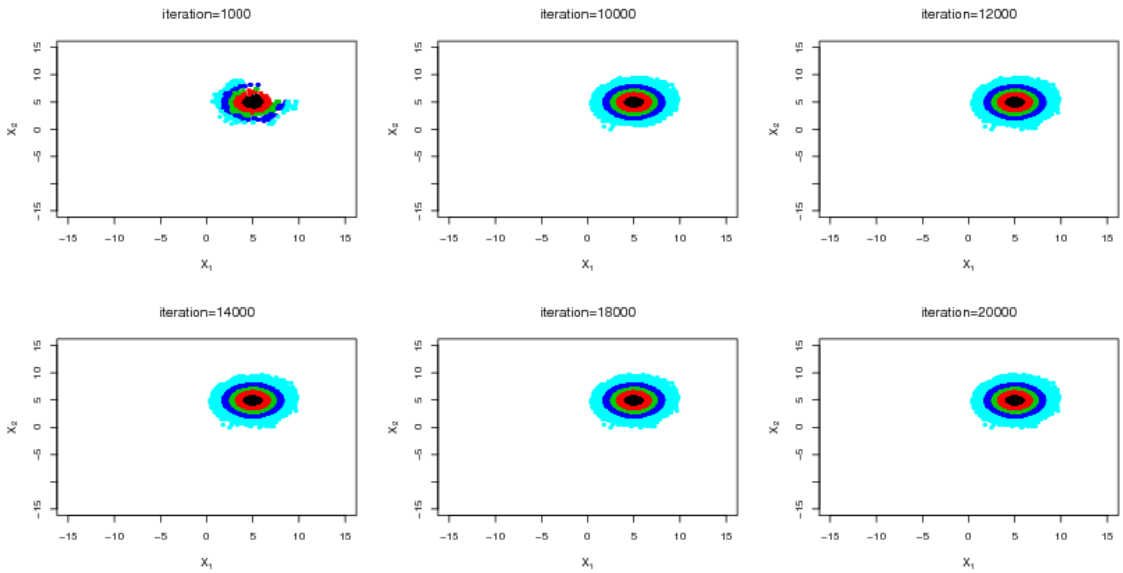
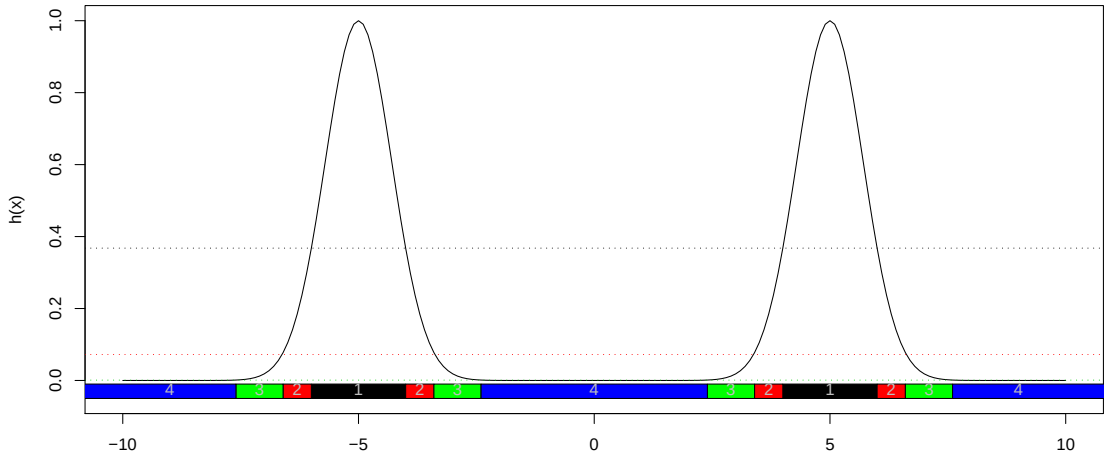


FIGURE 4.4: Sample path of GWL for the bimodal example using $E_{\max} = -\log(10^{-3})$. Although adaptive energy partitioning is applied to the internal energy levels, the chain still gets trapped due to under-estimation of maximum energy $E_{\max}$.

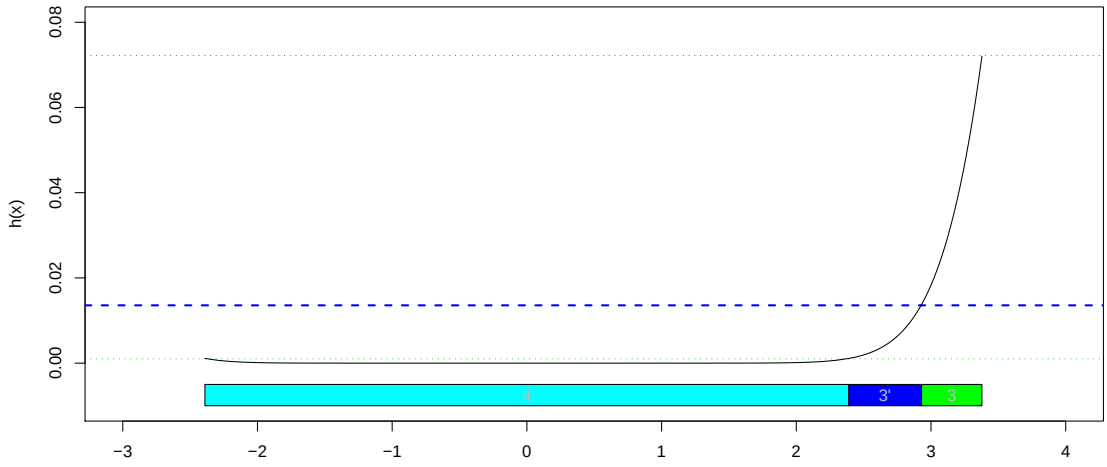
Figure 4.6 shows the results of applying the fully adaptive AE-GWL algorithm, which updates both internal and maximal energy levels. $E_{\max}$ is initially set to be $-\log(10^{-3})$, but this poor estimate is quickly overcome (compare Figures 4.3 and 4.6), and the energy barrier is crossed before $10,000$ iterations.

Note that in complex examples $E_{\max}$ can grow quite large, and the chain may spend increasing time in irrelevant regions of the state space, searching for yet another mode beyond ever higher energy barriers. Although this is not problematic, it may represent an efficiency issue for Monte Carlo integration with the resulting sample path (see Section 4.4), and it may be desirable to set some upper limit. (This is not the same as attempting to estimate $E_{\max}$ in the initialization of the energy partition. Instead, it gives a maximum barrier height beyond which the user is willing to assume no such barriers exist in the problem, as the algorithm will not attempt to cross regions of lower density.)

Notice that the target distributions in the above discussion are continuous. For



(a) Similar to Figure 4.2, but showing energy bins partitioned by the geometric progression determined by $E_{\max} = -\log(10^{-3})$. Bin 4 is too wide compared with the random walk scale 1, making it difficult for the chain to escape to the left.



(b) Closer view of the target density over the subregions $\mathcal{X}_3$, $\mathcal{X}_3'$, and $\mathcal{X}_4$, where energy bin $3'$ is split from the initial energy bin 3. Comparing with Figure 4.4 and Figure 4.5a, we see that although internal energy partitioning enables the chain to arrive in energy bin 4 from bin 3 through new bin $3'$, the chain remains unable to cross the energy barrier at 0 to discover the second mode, as bin 4 is too wide and suffers the same exponential decay of density as described previously.

FIGURE 4.5: Energy partition for the bimodal example using larger $E_{\max}$.

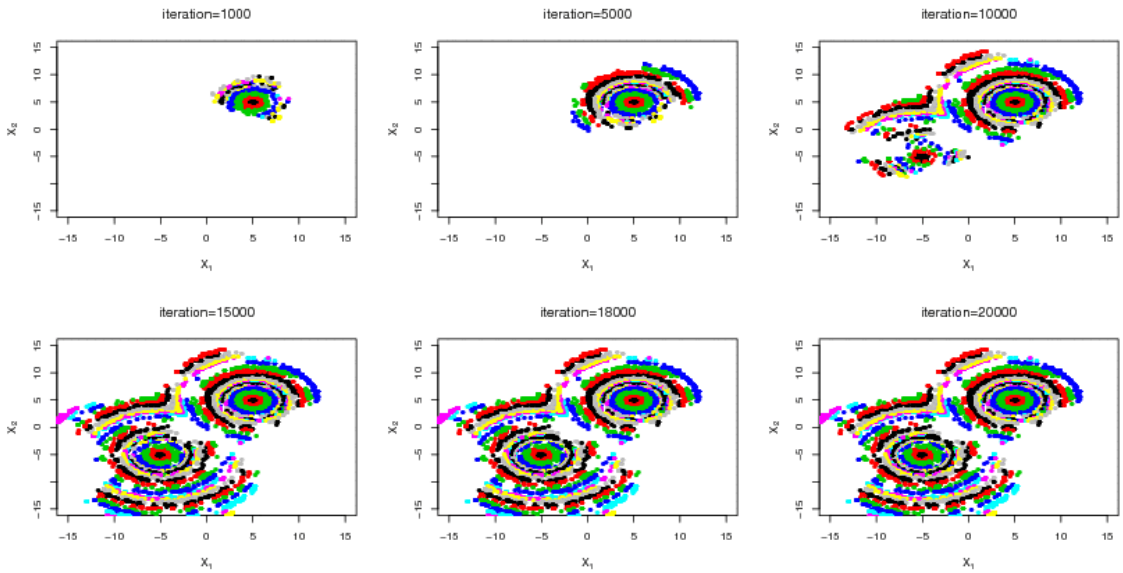


FIGURE 4.6: Sample path of AE-GWL algorithm on the bimodal example by using $E_{\max} = -\log(10^{-3})$ initially. The chain escapes from the initial mode $(5, 5)$, crossing the energy barriers to the other mode in less than $10,000$ iterations. Nine additional energy rings (colors) can be seen which were added by automatic partition refinement, enabling the crossing.

discrete problems, it can be impossible to enter some energy bins when no such states exist. To address this, we can add one more additional step to the algorithm for problems with discrete states. The step considers removal of unnecessary energy levels and will be used in the Ising model discussed in Section 4.5.

Algorithm 4.3.3. *Add to Algorithm 4.3.2:*

4. *For every n_{remove} iterations, update the energy partition by checking for unnecessary energy levels: if ϕ_{i-1} and ϕ_{i+1} are both positive, while $\phi_i = 0$, remove energy level E_i and modify d and ϕ accordingly.*

Up to this point, the geometric mean is used for determining the location of the new energy bin boundaries. However, this is not the only possible choice. In the remaining of this section, we will consider the conditions under which various choices may be optimal, and explore robustness to the choice of the splitting rule.

Consider the partition along the axis between two modes, as in Figure 4.2. Suppose the boundaries of the ith energy bin are $\mathcal{X}(E_i)$ and $\mathcal{X}(E_{i+1})$, and that the marginal density $\pi(x)$ or $h(x)$ decays over the interval from $\mathcal{X}(E_i)$ to $\mathcal{X}(E_{i+1})$. Let $L_i = \mathcal{X}(E_{i+1}) - \mathcal{X}(E_i)$ denote the width of the bin, and suppose we have decided to split this bin. We wish to choose a split such that the *sum* the expected stopping times in the two new energy bins is as small as possible. Denoting by $\tau_{A_\eta}^{(i)*}$ and $\tau_{A_\eta}^{(i+1)*}$ the expected stopping times in the two new energy bins formed by adding a new energy level E_{i+1}^*, we wish to choose E_{i+1}^* such that $\tau = \tau_{A_\eta}^{(i)*} + \tau_{A_\eta}^{(i+1)*}$ is minimized.

Let $L_{i*} = \mathcal{X}(E_{i+1}^*) - \mathcal{X}(E_i)$ and $L_{(i+1)*} = \mathcal{X}(E_{i+1}) - \mathcal{X}(E_{i+1}^*)$ be the new widths. Since $L_{i*} + L_{(i+1)*} = L_i$ and $\tau_{A_\eta}^{(j)*} \approx C_1 C_2^{L_{j*}/\eta}$ for some constants C_1 and C_2, τ achieves the minimum when $L_{i*} = L_{(i+1)*}$, and the optimal E_{i+1}^* is obtained by solving

$$2\mathcal{X}(E_{i+1}^*) = \mathcal{X}(E_i) + \mathcal{X}(E_{i+1}).$$

Not surprisingly, this depends on the form of the target density. If $\pi(x) \propto e^{-x}$, we have $\mathcal{X}(E_i) = E_i$, so $E_{i+1}^* = (E_i + E_{i+1})/2$ is just the arithmetic average. For the bimodal example used in this chapter, we have $\pi(x) \propto e^{-x^2/2}$, then $\mathcal{X}(E_i) = \sqrt{2E_i}$, giving

$$E_{i+1}^* = \frac{1}{2}\left(\frac{E_i + E_{i+1}}{2} + \sqrt{E_i E_{i+1}}\right),$$

an equal combination of the arithmetic and geometric means. When $\pi(x) \propto e^{-e^x}$, we get $\mathcal{X}(E_i) = \log(E_i)$ and $E_{i+1}^* = \sqrt{E_i E_{i+1}}$, suggesting that the geometric splitting rule is optimal when the density decays faster than a polynomial in x.

We perform simulations using the bimodal target distribution to explore robustness to suboptimal choices of splitting rule. Each test used the same random seed and ran three chains each for one splitting rule. In all simulations, the initial energy

levels were set up by $d = 4$ and $E_{\max} = -\log(10^{-3})$. Both internal energy levels and E_{d-1} were divided as needed.

We have shown that splitting energy bins is necessary in order to make the chain cross the energy barriers. However, the frequency of checking and splitting (determined by parameters n_{split} and threshold value E_{th}) does not really matter much to the final result in terms of whether it can cross and reach the other modes, especially for long chains. In some cases, more (extra) splits might speed up crossing. Therefore, in our simulations, we not only compare splitting rules for the same parameters, but also against different threshold values across the rules to see which rule-threshold combination works more robust and efficient.

Figure 4.7 compares three rules for two threshold values. The three rules are arithmetic mean, average of arithmetic and geometric mean, and geometric mean, respectively. The two threshold values are 50 and 5 for E_{th}. We ran 30 simulations for each of the three rules with the same parameter $n_{\text{split}} = 200$ and plot the histograms of the number of iterations it took for the chain, starting from one mode, to reach the other mode.

We have particularly ran more simulations to compare the first (arithmetic mean) and third (geometric mean) rules for the setting $n_{split} = 200$ and $E_{th} = 5$. The results of 100 simulations for each rule were summarized with histograms in Figure 4.8.

4.4 Computing expectations

Since the output samples $X_1, X_2, \ldots X_n$ from the WL chain are drawn from a non-stationary sequence of distributions obtained by reweighing densities over subspaces, we cannot use direct sample averages to approximate expectations under the target distribution. The following procedures are applied to obtain an approximation of the target. One of the main theoretical results in (Atchade and Liu, 2010) is to show

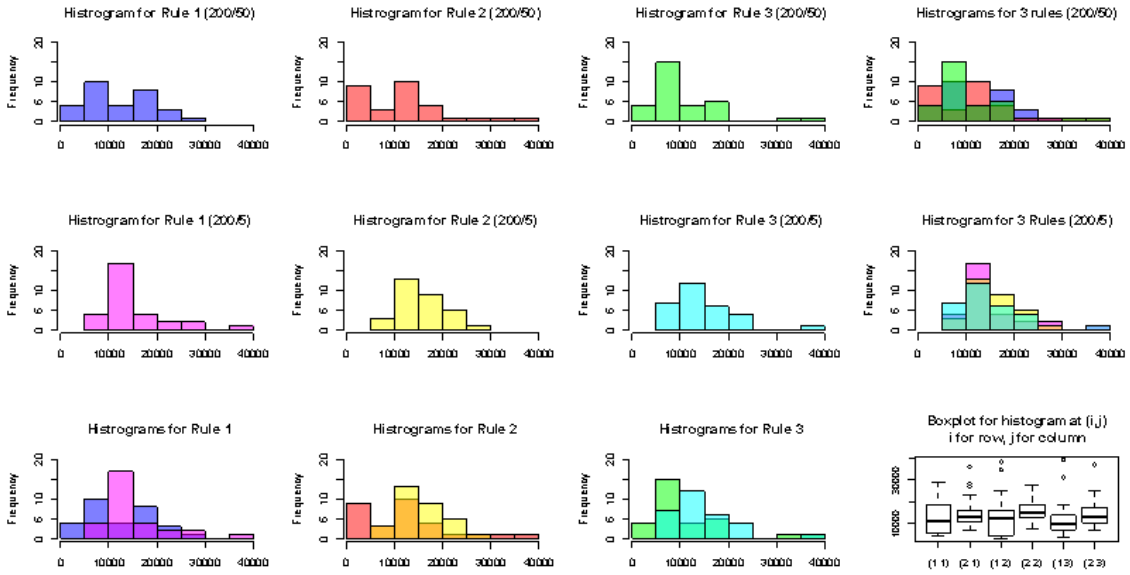


FIGURE 4.7: Comparison of crossing time for three splitting rules: arithmetic mean, average of arithmetic and geometric means, and geometric means. Histograms in the first two rows from the top, first three columns from the left summarize the results for six different rule-parameter combinations. The overlapped histograms in the third row from the top, and the forth column from the left were plotted to compare the histograms in the same row (column).

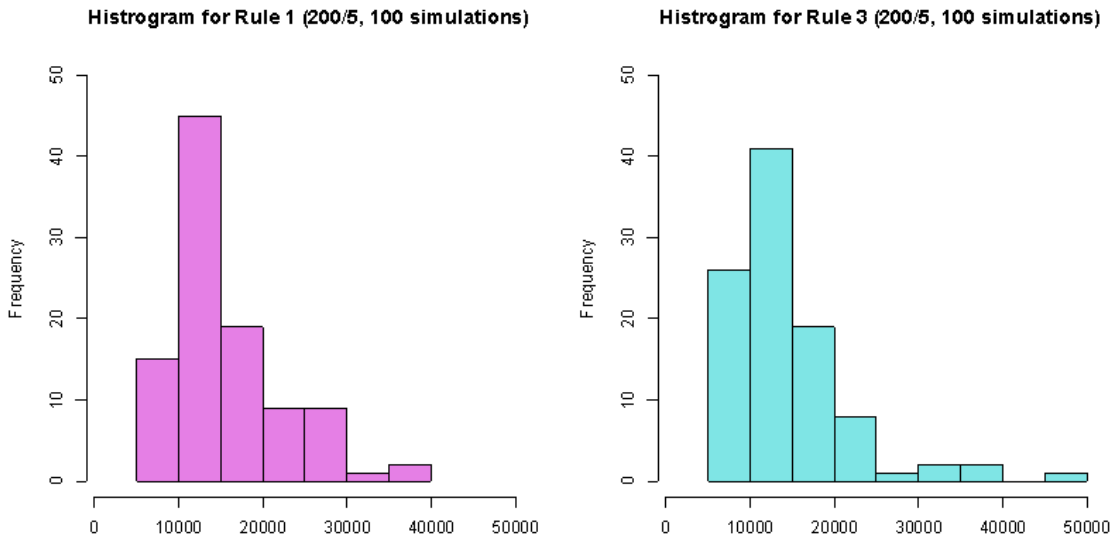


FIGURE 4.8: Comparison of crossing time for the first (arithmetic mean) and third (geometric mean) rules with 100 simulations.

that, for measurable functions f and any $i \in \{1, \ldots, d\}$ we have

$$\frac{1}{n_i} \sum_{k=1}^{n} f(X_k) \mathbf{1}_{\mathcal{X}_i}(X_k) \longrightarrow \int_{\mathcal{X}_i} f(x) \frac{\pi(dx)}{\phi(i)}, \quad \text{as } n \to \infty,$$

for any $i \in \{1, \ldots, d\}$, where $n_i = \sum_{k=1}^{n} \mathbf{1}_{\mathcal{X}_i}(X_k)$. However, in practice we find the rate of convergence of this estimator to be quite slow, perhaps unsurprisingly given the plug-in estimates in the denominator of the density. We address this by instead using an importance resampling scheme, as follows:

1. Use all samples $X_1, \ldots, X_n$ to form a kernel density estimate $\hat{f}$.

2. Apply importance resampling by resampling x_i, $i = 1, \ldots, m$ from $X_1, \ldots, X_n$ with weights $w_i = \frac{h(x_i)}{\hat{f}(x_i)}$.

The procedure is straightforward, and turns the GWL samples into samples from the target distribution of interest (Figure 4.9). Figure 4.10 compares the estimate for $E_\pi(X)$ in the first energy level obtained from the Atchade-Liu result with our importance resampling estimate, in terms of absolute error from the true value. Although the error in Atchade and Liu's method converges to 0 in the limit, our importance resampling approach converges significantly faster, essentially immediately after the chain crosses the energy barrier.

4.5 Results

We demonstrate the performance the AE-GWL algorithm on several multimodal target distribution examples.

4.5.1 Mixture distributions

We begins with a tridmodal mixture distribution

$$\pi(\mathbf{x}) = \frac{1}{3} \left[N(\mathbf{x}; (-3, -3)^T, I) + N(\mathbf{x}; (7, 7)^T, I) + N(\mathbf{x}; (5, -5)^T, I) \right],$$

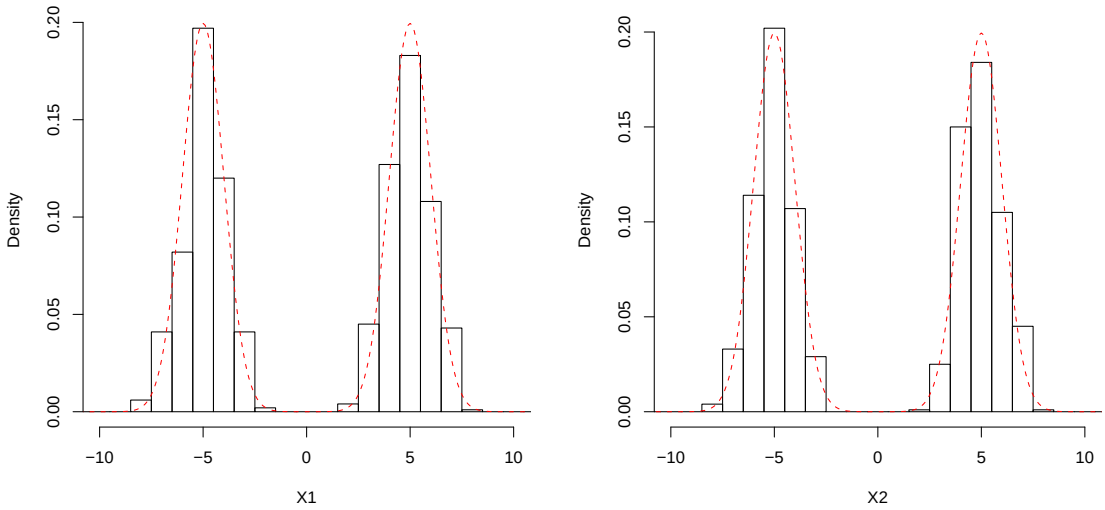


FIGURE 4.9: Histogram of $2,000$ samples obtained by importance resampling procedure applied to $10,000$ iteration AE-GWL sampling run. Red dotted line: true density. Resampling effectively produces samples with the target distribution of interest.

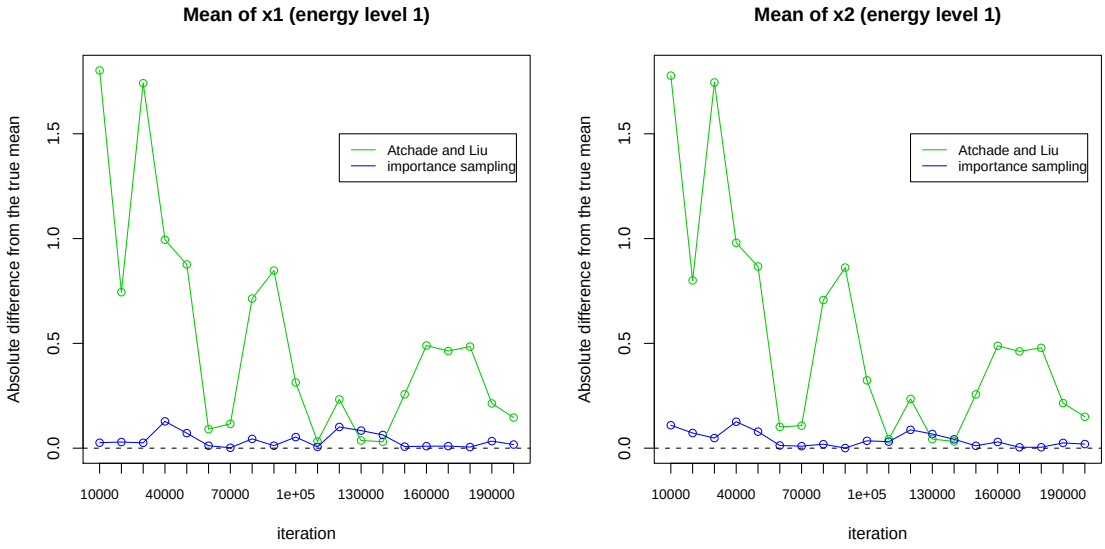


FIGURE 4.10: Absolute errors of the estimated $E_\pi(X)$ in the first energy level according to Atchade-Liu estimator (green) compared with our importance resampling scheme (blue). Convergence to zero is significantly faster for importance resampling.

which represents a rougher energy landscape in two dimensions than the two component mixture used earlier. Figure 4.11 shows the performance of the AE-GWL algorithm on this target distribution, including the sample path of the chain (Figure 4.11a) and estimated marginal densities (Figure 4.11b). Both barriers are successfully crossed.

Figure 4.12 shows the performance on a bimodal distribution in three dimensions

$$\pi(\mathbf{x}) = \frac{1}{2}\left[N(\mathbf{x}; (-4,-4,-4)^T, I) + N(\mathbf{x}; (4,4,4)^T, I)\right].$$

Again, the barrier is easily crossed and the marginal densities well-approximated.

4.5.2 Ising model

We next consider a two-dimensional Ising model (or spatial autologistic model) on an $L \times L$ 2D square lattice:

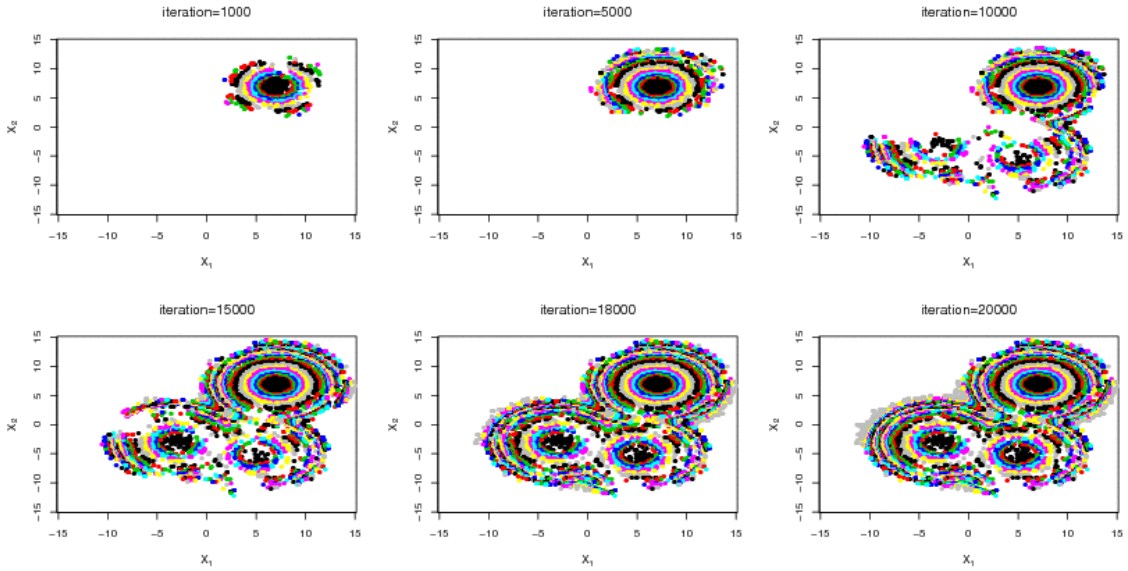
$$\pi(x) \propto h(x) = \exp\left\{\alpha \sum x_i + \theta \sum_{i\sim j} 1_{\{x_i = x_j\}}\right\}, \quad x_i \in \{0,1\}, \quad \alpha > 0, \quad \theta \in \{5,-5\}.$$

For $\theta > 0$, we can initialize all energy levels to be negative since $h(x)$ is positive. In addition, we can immediately set $E_{\min} = -\alpha L^2 - 2\theta L(L-1)$ and $E_{\max} = -\alpha\lfloor L^2/2 \rfloor$. Let F_1 and F_0 represent the set of states where the fraction of 1's is greater than or less than 0.5, respectively. Then

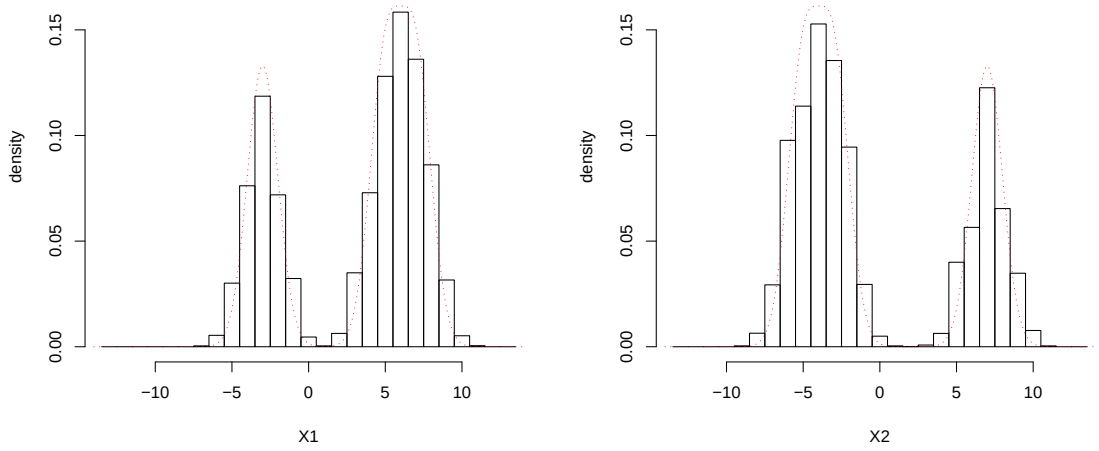
$$\frac{\sum_{x\in F_1}\pi(x)}{\sum_{x\in F_0}\pi(x)} = \frac{\sum_{x\in F_1}h(x)}{\sum_{x\in F_0}h(x)} = \frac{\sum_{x\in F_1}\exp\left\{\alpha\sum x_i + \theta\sum_{i\sim j}1_{\{x_i=x_j\}}\right\}}{\sum_{x\in F_0}\exp\left\{\alpha\sum x_i + \theta\sum_{i\sim j}1_{\{x_i=x_j\}}\right\}}$$

$$\approx \frac{\sum_{x_i=1,\forall i}\exp\left\{\alpha\sum x_i + \theta\sum_{i\sim j}1_{\{x_i=x_j\}}\right\}}{\sum_{x_i=0,\forall i}\exp\left\{\alpha\sum x_i + \theta\sum_{i\sim j}1_{\{x_i=x_j\}}\right\}} = \frac{\exp\{\alpha L^2 + 2\theta L(L-1)\}}{\exp\{2\theta L(L-1)\}} = e^{\alpha L^2},$$

where the approximation is due to the fact that the density of the states of all 1's and 0's dominates the other terms in the denominator and numerator. (The proportions



(a) Sample path at 20,000 iterations



(b) Density estimate for 20,000 iterations, obtained from 10,000 importance resamples. Histograms match the true density (red dotted lines) very accurately.

FIGURE 4.11: Performance of AE-GWL algorithm on trimodal target distribution.

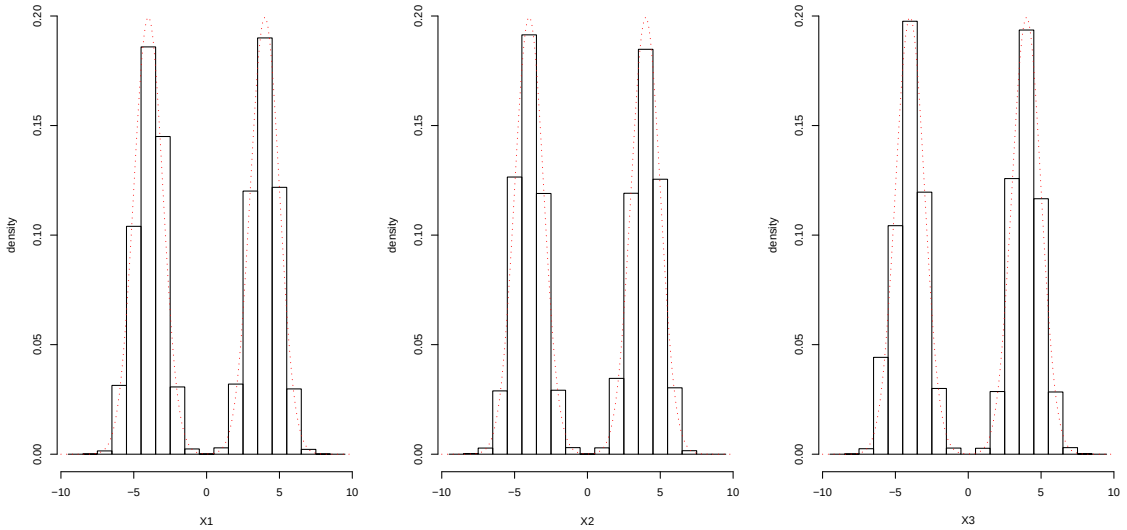


FIGURE 4.12: Three dimensional mixture distribution. Histograms obtained from $30,000$ iterations of AE-GWL algorithm, followed by $10,000$ importance resamples. Red lines (true density) are well approximated.

of the largest and second largest terms in the sum are $\exp\{2\theta+\alpha\}$ and $\exp\{2\theta-\alpha\}$, respectively. When θ is large relative to α, say $\theta = 5$ and $\alpha = 1$, the proportions are large enough for the other terms to be dropped in the summation.)

The above proportion $\sum_{x\in F_1}\pi(x)/\sum_{x\in F_0}\pi(x)$ says that even though both states of all 1's and all 0's are local modes, there is a significant preference on the states with more 1's than 0's, especially when L is large.

Taking $d = 20$, we set the initial energy levels by

$$E_0 = E_{\min}, \quad \ldots, \quad E_{i+1} = E_i + r_e^{i+1}, \quad \ldots, \quad E_d = E_{\max},$$

so that the differences between adjacent energies follow a geometric progression. We initialize the chain in the state of all 0's. Figure 4.13a presents results of AE-WL algorithm with updating energy levels for $\alpha = 1$, $L \in \{10, 15, 20, 30\}$. The chains escape the initial mode, cross the energy barrier and reach the other mode successfully.

For $\theta < 0$, the energy levels can be both positive and negative. We set $E_{\min} =$

$-\alpha\lceil L^2/2\rceil$ and $E_{\max} = 2\theta L(L-1)$. To initialize the energy levels, we can combine the strategies used in the previous examples for cases where all energy levels are positive or negative. We initialize the chain in one of the checkerboard state and compute $1/L^2$ of the Hamming distance from the initial state. Figure 4.13b shows the AE-WL chains reaches the other checkerboard for $L \in \{10, 15, 20, 30\}$.

4.5.3 Bayesian analysis of mixture exponential regression model

Suppose that

$$y_i \sim \begin{cases} \text{Exp}[\theta_1(\mathbf{x}_i)] & \text{with probability } \alpha = 0.3, \\ \text{Exp}[\theta_2(\mathbf{x}_i)] & \text{with probability } 1-\alpha = 0.7, \end{cases}$$

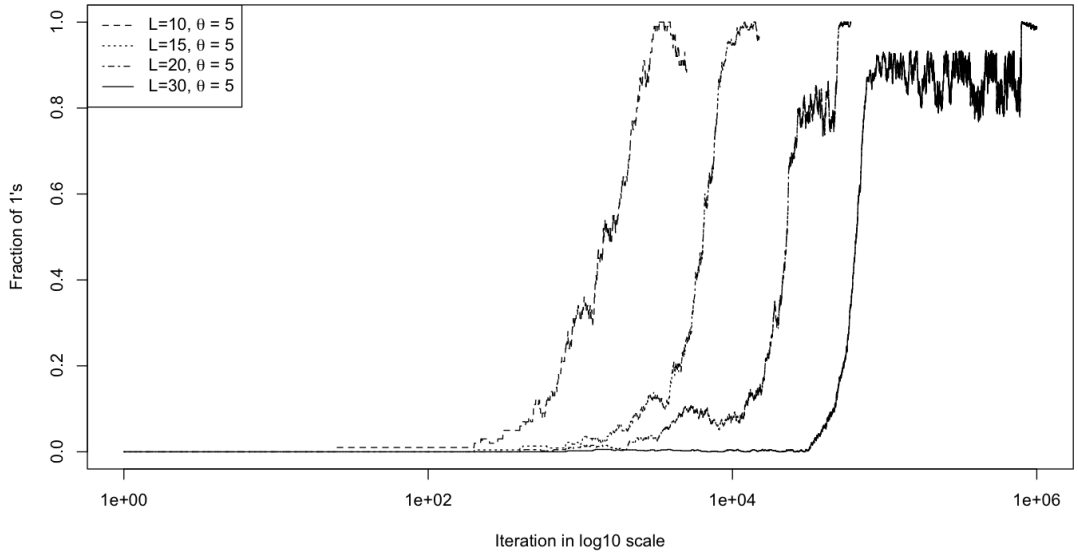
where $\theta_j(\mathbf{x}_i) = \exp(\boldsymbol{\beta}_j^T \mathbf{x}_i)$ for $j \in \{1, 2\}$, $\boldsymbol{\beta}_1 = (1, 2)^T$, $\boldsymbol{\beta}_2 = (4, 5)^T$, and $\mathbf{x}_i = (1, u_i)^T$ for $i = 1, \ldots, n$, with u_i's independently drawn from Unif$(0, 2)$. We wish to infer parameters α, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ based on observed data $\{\mathbf{x}_i, y_i\}_{i=1}^n$. The likelihood is of the form:

$$L(Y|\alpha, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \propto \prod_{i=1}^{n} \left[\frac{\alpha}{\theta_1(\mathbf{x}_i)} \exp\left(-\frac{y_i}{\theta_1(\mathbf{x}_i)}\right) + \frac{1-\alpha}{\theta_2(\mathbf{x}_i)} \exp\left(-\frac{y_i}{\theta_2(\mathbf{x}_i)}\right) \right].$$

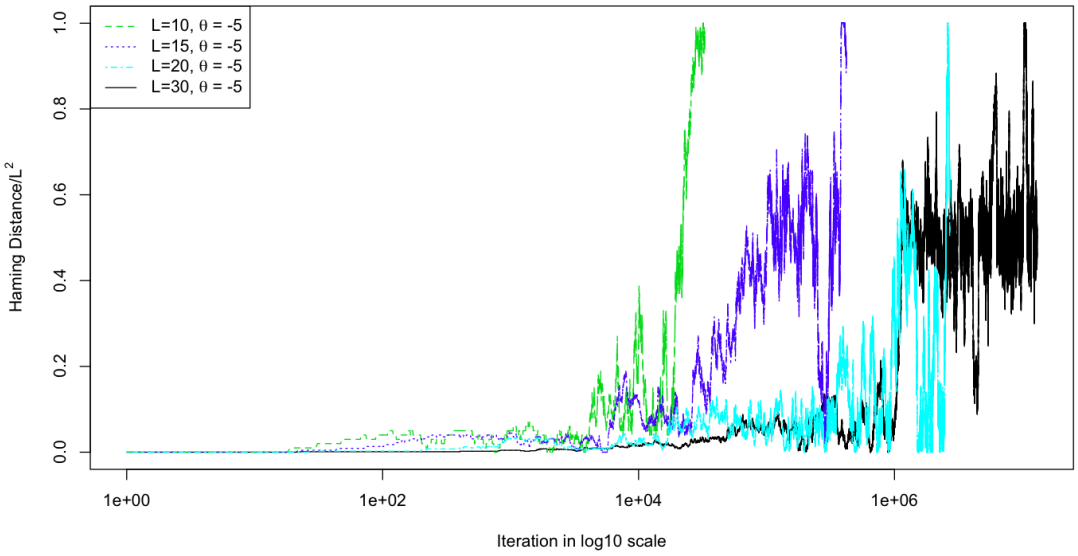
If we assign prior distributions $\pi_0(\alpha) =$Beta$(1, 1)$, $\pi_0(\boldsymbol{\beta}_j) =N(0, \sigma^2\mathbf{I})$ for $j = 1, 2$ and $\sigma = 10$, then the posterior distribution $\pi(\alpha, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2|Y)$ exhibits strong bimodality due to a "label switching" problem. It therefore provides an excellent model of a multimodal posterior distribution on which a GWL algorithm would naturally be applied. The energy then takes the form:

$$E(\alpha, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = -\log(\pi(\alpha, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2|Y)) = -\ell(Y|\alpha, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2) + \frac{1}{2\sigma^2} \sum_{k=1}^{2} \sum_{j=1}^{2} \beta_{kj}^2 + C,$$

where function ℓ is the log likelihood function. We set up $d = 10$ initial energy levels, determined by a geometric progression between $E_0 = 1700$ and $E_{d-1} = 2000$, and



(a) Traceplots of the fraction of 1's ($\|X\|^2/L^2$) for $\theta = 5$.



(b) Traceplots of $1/L^2$ Hamming distance for $\theta = -5$. Each trajectory shows the AE-GWL chain successfully escape from the initial mode and reaches the other mode.

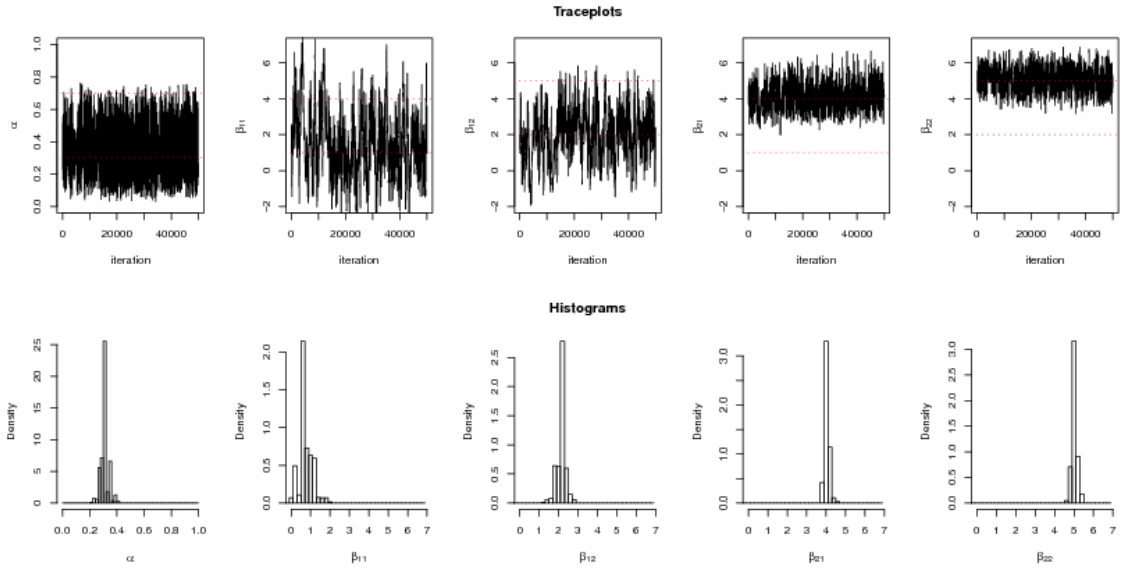
FIGURE 4.13: AE-GWL for Ising model on $L \times L$ 2D lattice, where $\alpha = 1$ and $L \in \{10, 15, 20, 30\}$.

applied the GWL and AE-GWL algorithms to this problem. The AE-GWL limit for $E_{\max}$ was set at 5000. Figure 4.14 compares the results. We see that without updating energy levels, the GWL chain fails to discover the symmetric mode, leading to a significantly incorrect posterior distribution and biased parameter estimates. In contrast, AE-GWL chain moves easily between the two modes to generate the correct posterior distribution.

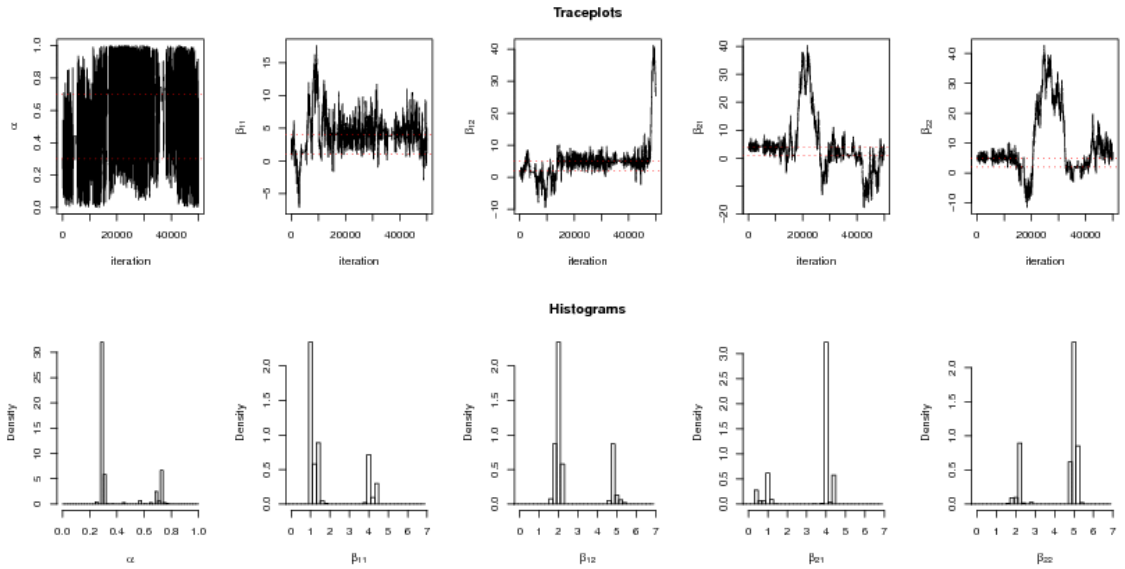
4.6 Discussion

In this chapter, we propose an adaptive energy partitioning scheme for the generalized Wang-Landau algorithm. The GWL algorithm provides an interesting approach to crossing energy barriers for sampling multimodal Bayesian posterior distributions general spaces; however, we have shown that its performance depends greatly on the predefined energy partitions and may fail even for simple low-dimensional bimodal distributions. We have also shown that this is due to fundamental restrictions on convergence imposed by the width of energy bins, and maximum energy height, specified by the initial partition. We have introduced an automatic, adaptive energy partitioning scheme that addresses these issues and performs well across a variety of examples in a fully automatic and user-independent way.

A remaining question for further research involves the effect of the choice of E_{th}—the threshold for deciding whether to subdivide energy bins—on the efficiency of the algorithm. In all the examples studied in this chapter, taking a default constant constant $E_{th} = 5$ worked well. However, if E_{th} is set too large one would expect longer time required between splits, and an adaptive selection of E_{th} may lead to further speedups. We are currently exploring such possibilities.



(a) The GWL chain gets trapped in the initial mode and never discovers the other half of the posterior distribution.



(b) The AE-GWL chain escapes the initial mode and generates samples from the correct posterior.

FIGURE 4.14: Posterior distributions and parameter traceplots for the mixture exponential regression model.

5

Adaptive Markov Chain Monte Carlo: An Exploration/Exploitation Approach

5.1 Introduction

In adaptive MCMC, the transition kernel of the chain is sequentially modified over time based on the current sample history. Common approaches include tuning proposal kernels for random-walk Metropolis chains and proposal distributions for Metropolized independence samplers (MIS) (Andrieu and Thoms, 2008; **?**). For instance, Ji and Schmidler (2013) provide a general approach to design MIS kernels by online minimization of the Kullback-Leibler divergence between the target distribution and proposal distributions, and show dramatic improvements in autocorrelation for Bayesian regression and variable selection problems. The recent equi-energy sampler (EES) (Kou et al., 2006) can similarly be viewed as building a (nonparametric) adaptive proposal for parallel tempering.

Recently, Schmidler and Woodard (2013) have shown that wide classes of adaptive MCMC algorithms including those mentioned above, fail to qualitatively improve convergence rates for multimodal distributions. Rather, they improve *autocorrela-*

tion for chains once equilibrium has reached. Other recent developments in MCMC algorithms can be viewed as a completely different type of adaptation. For example Atchade and Liu (2010) and Liang (2005) describe a continuous state-space generalization of the Wang-Landau algorithm of statistical physics. This algorithm (described in Section 5.2) partitions the state-space into subsets $\mathcal{X} = \cup_{i=1}^{d} \mathcal{X}_i$ according to energy (log-density), and adaptively estimates marginal probabilities of sets, in order to reweigh the target distribution $\pi(x)$ on each component to achieve uniform sampling across sets. Such approaches can be viewed as adaptive MCMC where the target distribution itself is adaptively modified, rather than the proposal. This is of great interest to us as it offers the possibility of circumventing/escaping the limitations identified by Schmidler and Woodard (2013).

In this chapter, we propose an "Exploration/Exploitation" (XX) approach to constructing adaptive MCMC algorithms, which combines adaptation schemes of distinct types. One piece, the "exploration" piece, uses adaptation strategies aimed exploring new regions of the target distribution and thus improving the rate of convergence to equilibrium. The second piece, the "exploitation" piece, involves an adaption component which decreases autocorrelation for sampling among regions already discovered. This hybrid combination is relatively simple, yet provides the best of both worlds. As an example of this approach, we develop an XX algorithm that combines an Adaptive Metropolized Independence Sampler (AMIS) as the exploitation component, with the generalized Wang-Landau (GWL) algorithm with the adaptive energy level partition scheme (Chapter 4) as the exploration component. We show that, for multimodal target distributions, the AMIS algorithm requires general purpose modifications, which we provide. We demonstrate that the combined XX algorithm significantly outperforms either component algorithm on difficult multimodal sampling problems.

The organization of the chapter is the following. Section 5.2 briefly summarizes

the AMIS algorithm used in the XX algorithm. Section 5.3 describes how to apply the AMIS algorithm in combination with the WL algorithm from Chapter 4 in the framework of XX. One particular simulation example reveals the limitation of the AMIS, which is addressed with several improvements in Section 5.4. We apply the revised AMIS algorithm and WL algorithm to a mixture regression problem and a neural network examples to demonstrate the performance of our improvements. We conclude this chapter in Section 5.6.

5.2 Adaptive Metropolized independence samplers

Adaptive Metropolized Independence Samplers (AMIS) have been developed in parallel by several authors (Andrieu and Thoms, 2008; Ji and Schmidler, 2013; Craiu et al., 2009). The general theme is the adaptive construction of a Metropolized Independence Sampler (MIS) proposal distribution to fit the current sample history, with the proposal distribution converging to an approximation of the target distribution in the limit. In this section, we summarize some relevant results of an AMIS scheme with mixture proposal distribution developed by Ji and Schmidler (2013), for the completeness of discussions in the rest of this chapter.

Let $\pi(x)$ and $q(x)$ denote the target distribution and the proposal distribution in an AMIS scheme, respectively. When $\pi(x)$ is multimodal, using an unimodal $q(x)$ generally performs poorly, due to difficulty in approximating the posterior. An alternative approach is to consider $q(x)$ as a mixture distribution of the form:

$$q(x) = \lambda q_0(x; \tilde{\psi}) + (1 - \lambda) \sum_{m=1}^{M} w_m q_m(x; \psi_m),$$

where notation $q(x; \psi)$ denotes an proposal density with parameter ψ, λ and $\tilde{\psi}$ are fixed, and w_m and ψ_m are updated from previous samples using an stochastic approximation optimization. When $q(x; \psi)$ is a normal distribution $N(x; \mu, \Sigma)$, the

adaptive strategy is as follows:

Algorithm 5.2.1 (AMIS with mixture proposal distribution)**.** *For component m, denote the values of w_m, μ_m, and Σ_m in the nth iteration by $w_{m,n}$, $\mu_{m,n}$, and $\Sigma_{m,n}$, respectively. Let $(\mathbf{w}_n, \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n) = \{w_{m,n}, \mu_{m,n}, \Sigma_{m,n}\}_{m=1}^M$. Initialize $(\mathbf{w}_1, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$. At iteration $n+1$:*

1. *Draw a new sample X_{n+1} by MIS with respect to the proposal distribution*

$$q_n(x) = \lambda N(x; \tilde{\mu}, \tilde{\Sigma}) + (1-\lambda) \sum_{m=1}^{M} w_{m,n} N(x; \mu_{m,n}, \Sigma_{m,n}).$$

2. *Update the parameters $(\mathbf{w}_{n+1}, \boldsymbol{\mu}_{n+1}, \boldsymbol{\Sigma}_{n+1})$ by*

$$\begin{aligned} w_{i,n+1} &= w_{i,n} + r_{n+1}\left[O_i(X_{n+1}) - \bar{O}\right], \\ \mu_{i,n+1} &= \mu_{i,n} + \alpha_{i,n+1}(X_{n+1} - \mu_{i,n}), \\ \Sigma_{i,n+1} &= \Sigma_{i,n} + \alpha_{i,n+1}\left[(X_{n+1} - \mu_{i,n})(X_{n+1} - \mu_{i,n})^T - \Sigma_{i,n}\right], \end{aligned}$$

where $\alpha_{i,n+1} = r_{n+1} w_{i,n} O_i(X_{n+1})$,

$$O_i(X_{n+1}) = \frac{N(X_{n+1}; \mu_{i,n}, \Sigma_{i,n})}{\sum_{m=1}^{M} w_{m,n} N(X_{n+1}; \mu_{m,n}, \Sigma_{m,n})}, \quad \bar{O} = \frac{1}{M} \sum_{m=1}^{M} O_m(X_{n+1}),$$

and r_{n+1} is the step-size of the stochastic approximation algorithm (Robbins and Monro, 1951; Kushner and Yin, 1997).

For sufficiently large M, $q(x)$ can adapt to approximate $\pi(x)$ arbitrarily well, and therefore in the limit the AMIS chain will behave like independent sampling from π. In practice, the choice of M has been heuristic. In Section 5.4, we give a new, simple and effective approach for automatic determination of M.

Lastly, we comment here that in Algorithm 5.2.1, we have $\sum_m w_m = 1$ but not $w_m \geqslant 0$. Rather than add slack variables to satisfy the Karush-Kuhn-Tucker conditions, one can project back onto the unit simplex if the weights become negative, as

common in stochastic approximation. In the context of this chapter, for computational reasons, we define a lower bound for the weights, denoted by w_{lb}, and set it to be 0.01.

5.3 An exploration/exploitation algorithm for adaptive Markov chain Monte Carlo

Schmidler and Woodard (2013) suggest limitations of the AMIS method and related. Also, Wang-Landau (WL) algorithm has limitations in autocorrelation. These suggest combining the two. A natural approach is to consider a transition of the two, e.g., $K = \alpha K_{\text{AMIS}} + (1-\alpha)K_{\text{WL}}$. However, Schmidler (2012) shows that such kernels suffer from significant limitations or behave poorly for multimodal target distributions, despite the presence of the WL component. Instead, we propose an alternative way to combine these two algorithms in the hope that it will help the adaptive MIS to explore new regions in the state space. We refer to this algorithm as AMIS+WL, or XX.

Algorithm 5.3.1 (XX)**.** *Start with two chains, X^{WL} and X^{AMIS}, separately.*

1. *Every N_c iterations, update the proposal distribution for X^{AMIS} using samples from X^{WL} in two steps. First, obtain N_{wl} samples from all previous samples of X^{WL} by importance sampling method. Second, update w_m, μ_m and Σ_m with these N_{wl} samples.*

2. *Run two chains independently at other iterations.*

To present the performance of this algorithm, we consider a simple trimodal target distribution obtained by a mixture of normal distributions:

$$\pi(x) = \frac{1}{3}\left[N(x;(-8,-8)^T,I) + N(x;(8,8)^T,I) + N(x;(20,-20)^T,I)\right].$$

Table 5.1: Parameters for AMIS, WL, and XX algorithms for trimodal target distribution.

Algorithm	Parameters
AMIS	$\lambda = 0.01$, $\tilde{\mu} = (0,0))^T$, $\tilde{\Sigma} = 2I$, $M = 10$ $w_{i,1} = 1/M$, $\mu_{i,1} \sim N_2(\mathbf{0}, 2I)$, $\Sigma_{i,1} = I$, $r_n = 1/n$
WL	$d = 10$, $\gamma_n = 1/n$, $\epsilon = 0.3$, $E_{th} = 5$, $n_{\text{split}} = \text{nIter}/100$ upper bound on updting the highest energy level is 130
XX	$N_{wl} = 100$, $N_c = \text{nIter}/100$

This artificial distribution can be representative of many multimodal target distributions arising in Bayesian statistics, where posterior modes will be well approximated locally normality for adequate data sample sizes. We tested the AMIS, WL, and XX algorithms, with the parameters given in Table 5.1. All simulations were initialized at the sample point within mode $(8, 8)$. The AMIS chain remains in the mode where it was initialized. So did the AMIS component of the AMIS+WL chain, up until the first introduction of information from WL chain at iteration N_c.

The behaviors of the AMIS and WL chains are depicted in Figures 5.1 and 5.2, respectively. The first row in Figure 5.1 shows that the AMIS adapted proposal distribution at several time points. It is clear that the chain never escapes the initial mode. Unsurprisingly, the resulting marginal density approximations are poor. Furthermore, the autocorrelation plots are highly misleading, emphasizing the dangers of using single-chain plots for convergence diagnostics. In comparison, the first row in Figure 5.2 shows the sample path of the WL chain, which eventually crosses each energy barrier to find all three modes. Colors denote distinct energy partition levels. We observe that the WL chain starts to cross between the first and the second modes around iterations $30,000$–$40,000$, and from the second to the third modes around iteration $90,000$. The second row in Figure 5.2 shows that the autocorrelation is estimated to be very large, and demonstrates the density approximation obtained using the importance resampling scheme, described in Section 4.4.

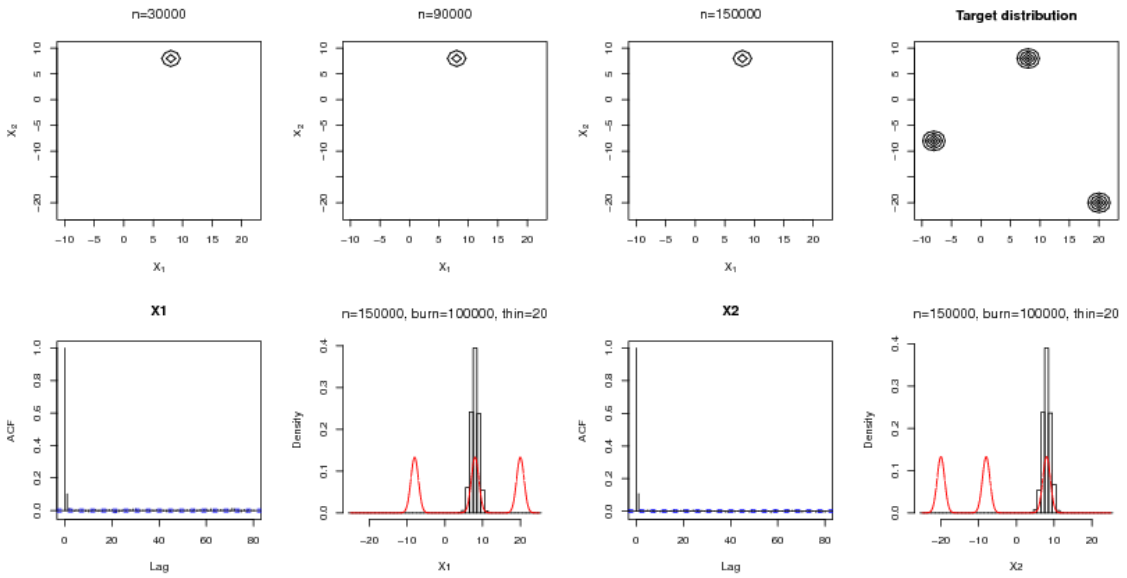


FIGURE 5.1: Simulation result of the AMIS algorithm for the trimodal target distribution. The contour plots in the first row shows that the mixture proposal of AMIS chain misses two of the three modes. The second row show the autocorrelation plots and marginal density approximations summarized from the samples of the AMIS chain. The density approximation is a poor match to the true marginal distributions (red curve).

The result of the XX algorithm is shown in Figure 5.3. The first row shows the evolution of the AMIS mixture proposal distribution when the WL sample information is included. In contrast to the standalone AMIS algorithm (Figure 5.1), we see the development of a new proposal component centered near $(0, 0)$ at approximately iteration $40,000$, immediately after the discovery of this region by the WL chain. However, although the WL chain reaches the third mode around $90,000$ iterations, the adapted AMIS proposal distribution does not reflect this new region, making the sampled distribution shown on the first two modes. This highlights a failure of the AMIS algorithm for multimodal target distributions which has not previously been appreciated, and is discussed in the next section. Notice that this failure of the AMIS was not observed before, due to its poor exploration ability. Furthermore, it suggests that in order to achieve the desired performance of the XX algorithm, the AMIS algorithm must be modified to enable reliable adaptation to multimodal

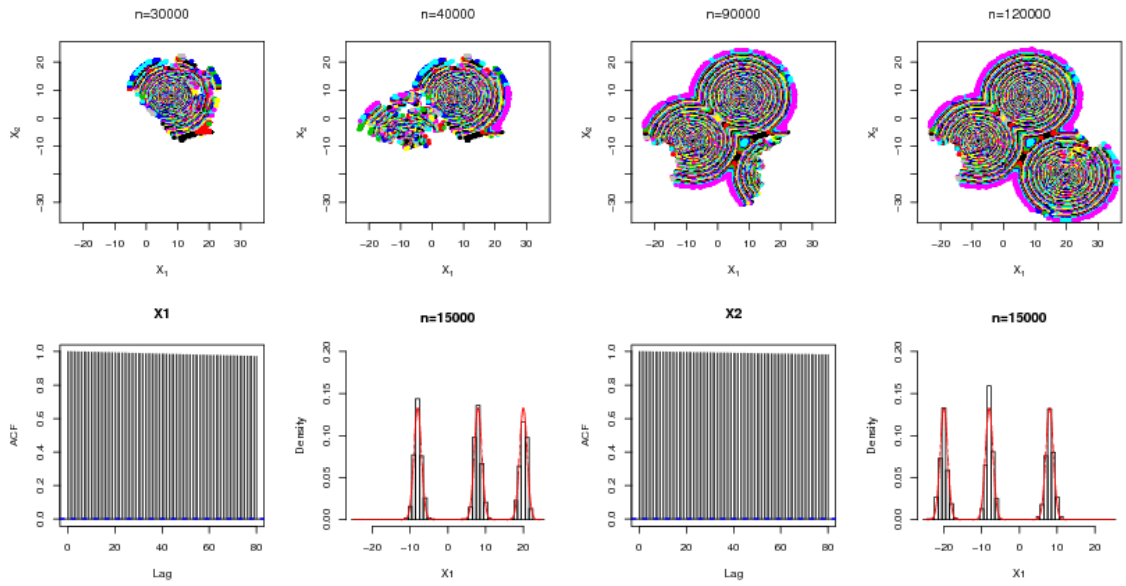


FIGURE 5.2: Simulation result of the WL algorithm for the trimodal target distribution. The sample path on the first row suggests that the WL chain starts at mode $(8, 8)$, and approaches the second mode $(-8, -8)$ around iteration $40,000$, and reaches the third mode $(20, -20)$ by iteration $90,000$. The second row shows autocorrelation and marginal density approximation results. The samples approximate the true marginal distributions well, but the autocorrelation is large with a very slow decay.

distributions. Section 5.4 describes such an extension, which can be used by itself or within the context of the XX algorithm.

5.4 Extension of the AMIS algorithm for adaptation to multimodal distributions

As described above, combining the AMIS algorithm with the WL algorithm to form the XX algorithm enables the chain to cross between modes which the AMIS algorithm alone is unable to reach. In doing so, it reveals a previously unobserved breakdown of the AMIS algorithm on multimodal distributions. (Although we describe this weakness in terms of the stochastic approximation formulation of the AMIS given by Ji and Schmidler (2013), it applies equally to the alternative (sequential EM update-based) AMIS formulation described by Andrieu and Thoms

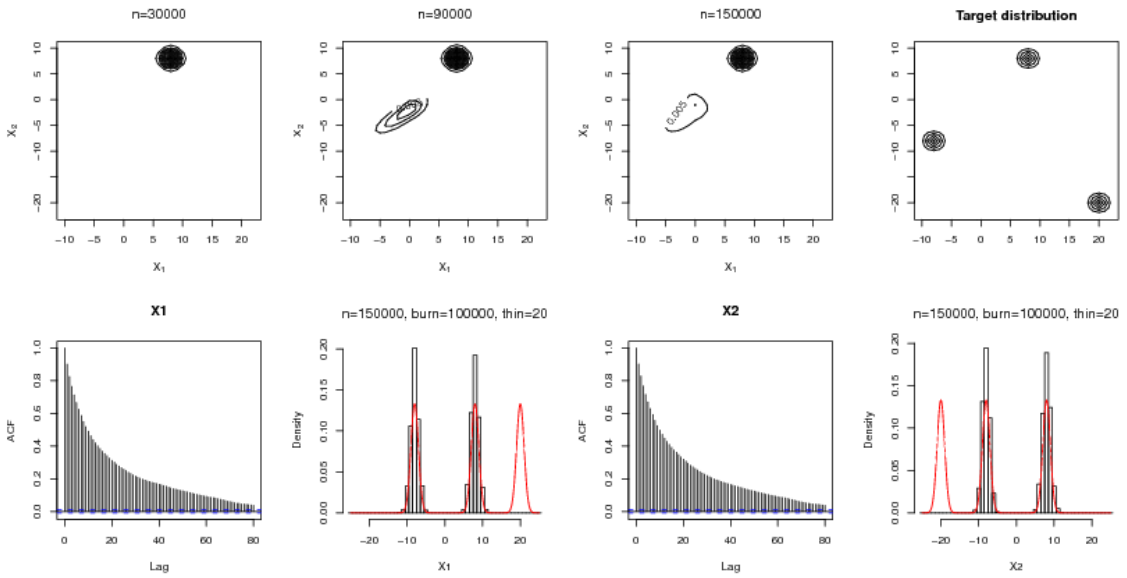


FIGURE 5.3: Simulation result of the XX algorithm for the trimodal target distribution. Contour plots in the first row implies two significant proposal components representing two of the three modes, missing the third one. The autocorrelation plots on the second row show quick decay of the autocorrelation as time lag between samples increases. The marginal density approximations show that the samples are able to capture two out of the three modes.

(2008), which also uses a decaying update weight). In particular, the sequential updating scheme of the AMIS mixture proposal does not accommodate very well large changes in the target distribution observed late in the sampling, which limits the ability to take full advantage of the information provided by the exploration chain in the context of the XX algorithm. For example, the decreasing step-size sequence r_n (required for convergence in the limit) means that, when the crossing time of WL is long, updates of proposal parameters are forced to be small and thus react very slowly. In addition and as a result, parameters may not adjust to approximate the target distribution accurately.

Consider the trimodal target distribution of the previous section. Figure 5.3 shows the failure of the AMIS algorithm to adapt to the third mode even when presented with samples from this mode (via the coupling to the WL chain). The mean of the proposal mixture component which approximates the second mode ap-

proaches the correct value $(-8, -8)$ between iterations $40,000$ and $90,000$, while the variance/covariance become larger and larger (since the new resampling sample X^* from WL chain is close to the second mode at $(-8, -8)$, but the mean is adapting slowly as r_n now is extremely small, $(X^* - \mu)(X^* - \mu)^T$ will be a matrix of large values. And the entries of Σ is increasing because every time it is updated, a positive value is added). (Figure 5.4). These adaptations occur at a slow rate because, by the time the second mode is found, the step-size sequence values r_n are quite small. When the WL chain discovers the third mode (approximately iteration 90k), large covariance of the second mode component along with the small r_n prevent the adaptation of any components to the third mode. While this problem can (might) be resolved by increasing the initial r_0 or slowing the rate of decay for r_n (while still satisfying $\sum_n r_n = \infty$ and $\sum_n r_n^2 < \infty$ to ensure convergence), the appropriate scale can only be known after all modes have been discovered (and the choices would be problem-specific); we cannot know in advance if a given r_n sequence is adequate. Instead, we introduce an algorithmic approach to automatically identify such situations and "reset" the r_n sequence as needed.

We make two modifications to the AMIS algorithm to improve the ability to adapt to multimodal distributions. The main idea is to recognize when "new" modes or regions of significant mass have been reached by the sampler, and allow for more rapid adaptation to resume temporarily. This requires three improvements on the standard AMIS updating scheme:

1. Identification of new (previously unvisited) modes or regions

2. Resetting step-size decay sequence to resume large adaptations

3. Automatic addition of new mixture components (increasing M) as needed

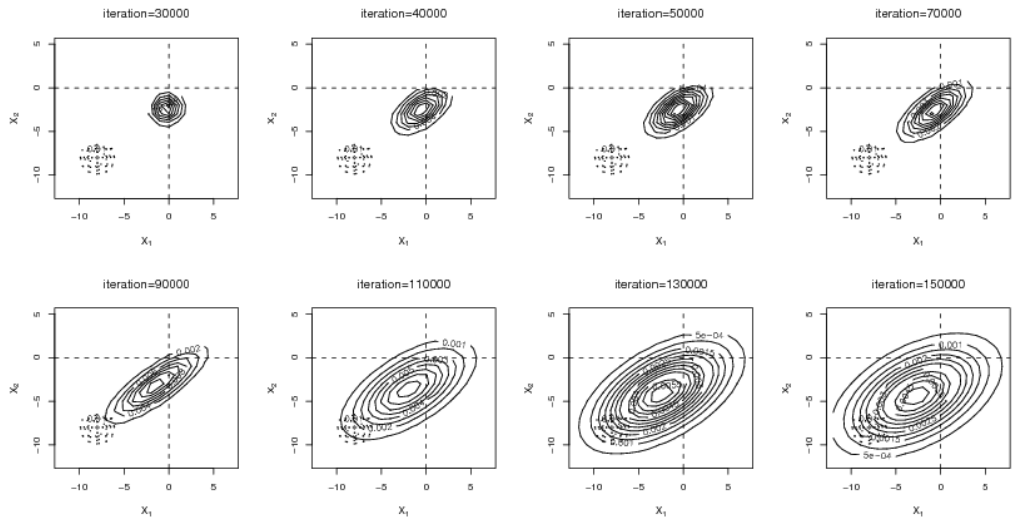


FIGURE 5.4: The dashed contour plot represents the second mode at $(-8, -8)$, and the solid contour plot is for the component (closest to that mode) of the mixture proposal distribution of XX chain. Although this proposal component gets closer to the second mode, both the mean and covariance are updated slowly due to the small value of r_n when n is large.

5.4.1 Identification of new regions

We begin with a strategy for determining that a sample X_{n_0} represents a new mode or region when the quantity $O(x)$ in Algorithm 5.2.1 satisfies the condition $O_i(X_{n_0}) > D_{th}$ for some threshold D_{th}.

To choose a reasonable value for the threshold D_{th}, we first notice that

$$O_i(X_{n_0}) = \frac{1}{w_{i,n_o-1} + \sum_{m=1,m\neq i}^{M} w_{m,n_0-1} \frac{N(X_{n_0};\mu_{m,n_0-1},\Sigma_{m,n_0-1})}{N(X_{n_0};\mu_{i,n_0-1},\Sigma_{i,n_0-1})}} \leqslant \frac{1}{w_{lb}}.$$

Next, $O_i(X_{n_0})$ depends on what region X_{n_0} is in and whether the ith proposal component is significant (the corresponding weight is at least $1/M$), which can be divided into three cases. (1) X_{n_0} finds a new region, and $(\mu_{i,n_0-1}, \Sigma_{i,n_0-1})$ represents the negligible proposal component that is closest to this region, then $w_{i,n_0-1} = w_{lb}$, and in the summation, either w_{m,n_0-1} or the fraction is extremely small. Therefore $O_i(X_{n_0})$ can get any closer to $1/w_{lb}$, depending on how far those found regions are

away to the new region. (2) X_{n_0} is in one of the visited regions, and $(\mu_{i,n_0-1}, \Sigma_{i,n_0-1})$ represents the significant proposal component which is closest to this region, then $w_{i,n_0-1} \geqslant 1/M$, so $O_i(X_{n_0}) < M$. (3) If X_{n_0} is in one of the visited regions, but $(\mu_{i,n_0-1}, \Sigma_{i,n_0-1})$ is not the closest proposal component, then at least one term in the summation would be larger than $1/M$, so $O_i(X_{n_0}) < M$. In conclusion, the threshold value D_{th} could be any number between M and $1/w_{lb}$. For the examples considered in this chapter, we have $M = 10$ and $w_{lb} = 0.01$. Therefore, we choose to set $D_{th} = 50$.

5.4.2 Improvement 1: reset the step-size sequence $\{r_n\}$

We introduced a new strategy for updating parameters of the AMIS algorithm, which allows the revised algorithm to work much better than the original version. However, as shown in Figure 5.3, using samples from WL chain to update proposal distribution of AMIS chain alone is insufficient for the proposal distribution to approximate the target distribution arbitrarily well. As the step-size sequence decreases, the change in the parameters gets smaller and smaller, even when there should be some important updates. This suggests that we should reset the step-size sequence when necessary. Specifically, when $O_i(X_{n_0}) > D_{th}$, in addition to Algorithm 5.3.1, we also do

$$r_{n_0} = r_{\text{restart}} = 1, \quad r_n = \frac{1}{n - n_0 + \frac{1}{r_{\text{restart}}}}, \quad n > n_0. \tag{5.1}$$

Algorithm 5.4.1. *When $O_i(X_{n_0}) > D_{th}$:*

1. *Update $\{w_j, \mu_j, \Sigma_j\}_{j=1}^M$ using the regular formula with importance sampling samples from WL chain.*

2. *Modifty the step-size sequence $\{r_n\}$.*

We apply Algorithm 5.4.1 to the trimodal target distribution. We see that in Figure 5.5, the combination chain generates samples from the target distribution

with relatively low autocorrelation compared to the results obtained from WL chain, however, the mixture proposal distribution seems not to approach the target distribution as fast and accurate as we want. Although not shown in the contour plot for iteration 150000, there is in fact one significan proposal component with the mean close to $(20, -20)$ and a covariance matrix whose entries have large absolute values. Both of the new significant proposal components (with means at $(8, -8)$ and $(20, -20)$ respectively) were gradually developed after resetting the r sequence, which was supposed to result in fast updates in the mixture proposal distribution. The reason for this is that the new components with means and covariance matrices not close to the true modes and small weights cannot get updated soon after the combination iteration, as proposed samples are rejected with large probabilities. The component representing the mode at $(-8, -8)$ has a major update when the third mode was introduced around iteration 92000, while the third new component needs several rounds of combination and update with WL chain. Nevertheless, Figure 5.6 describes the weights can adapt and approximate the true values eventually. We can see around iteration 100000, the step-size sequence needs to be reset for a couple of times because the weight for the proposal component decreases as the component parameters are initially very different from the true values yet gradually get updated. Lastly, Figure 5.7 shows that there are six times when the step-size sequence needs to be reset, while in fact there are actually only two new regions which cannot be discovered by the AMIS chain.

From the above simulation results, it seems that restarting the step-size could help with speeding up the updating of parameters, however now as fast as we want it to be. The obstacle in the speedy update lies in the updating scheme. Using WL samples to update proposal parameters in the regular way may cause a huge covariance matrix for the new dominant component at the iteration when a new region is detected. By the updating formula, μ_i can be adapted relatively fast after

the modification of r. However, if the new sample X is too far away from μ_i, Σ_i can be very large. Hence, we consider another way to improve Algorithm 5.3.1–create a new proposal component when a new region is discovered instead updating the existing proposal components.

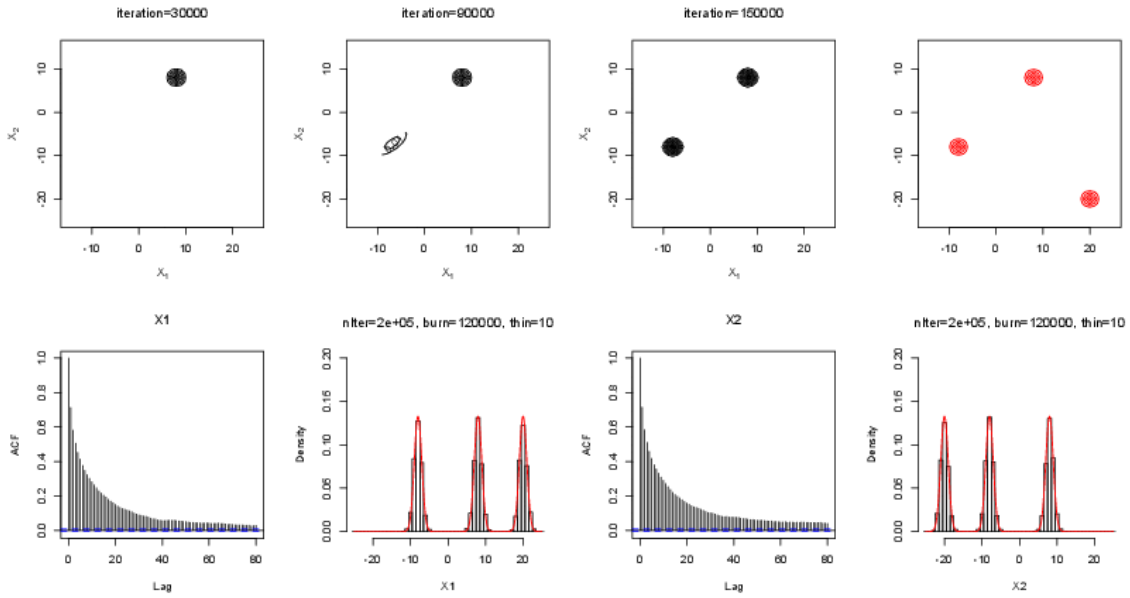


FIGURE 5.5: Simulation results of the XX algorithm using revised AMIS (Algorithm 5.4.1) for the trimodal target distribution. The marginal density approximations agrees with the target distribution, and the low autocorrelation imply that the mixing is good, even though the proposal distribution does not approximate the target very well.

5.4.3 Improvement 2: add new proposal components.

Figure 5.4 suggests that rapid adaptation cannot be realized by Algorithm 5.3.1, since the proposal component was initally far away from the newly discovered mode and small r cannot help with fast updating. In the previous section, we show results from the first improved algorithm by resetting step-size sequence and discuss its advantage and drawbacks. In this section, we propose another algorithm aiming to overcome the bad initialization.

Algorithm 5.4.2. *When* $O_i(X_{n_0}) > D_{th}$*:*

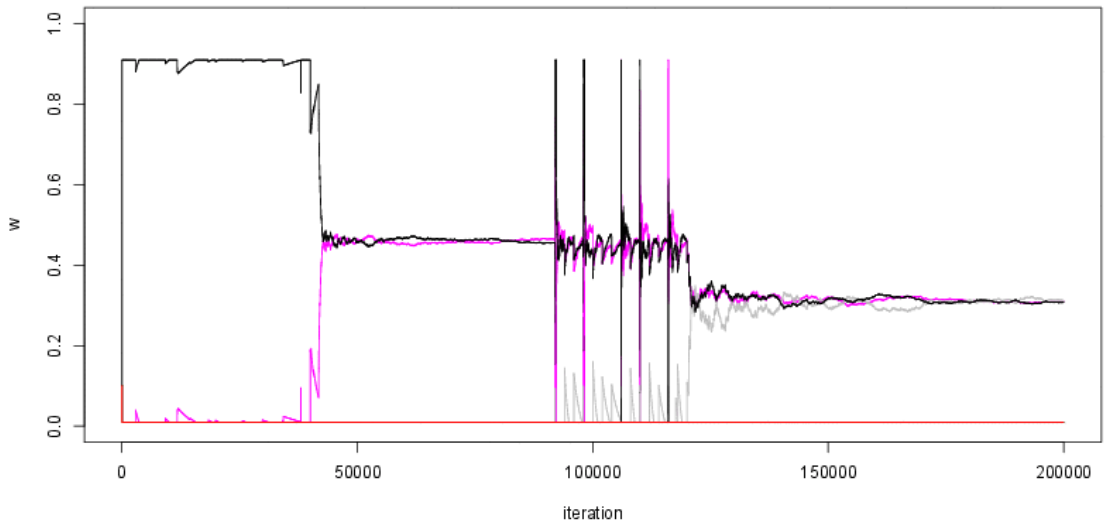


FIGURE 5.6: Weight sequence for XX using revised AMIS (Algorithm 5.4.2). Compared with the result in Figure 5.6 for XX using Algorithm 5.4.1, we can see that the weights for the new dominant components adapt much faster to the true proportions.

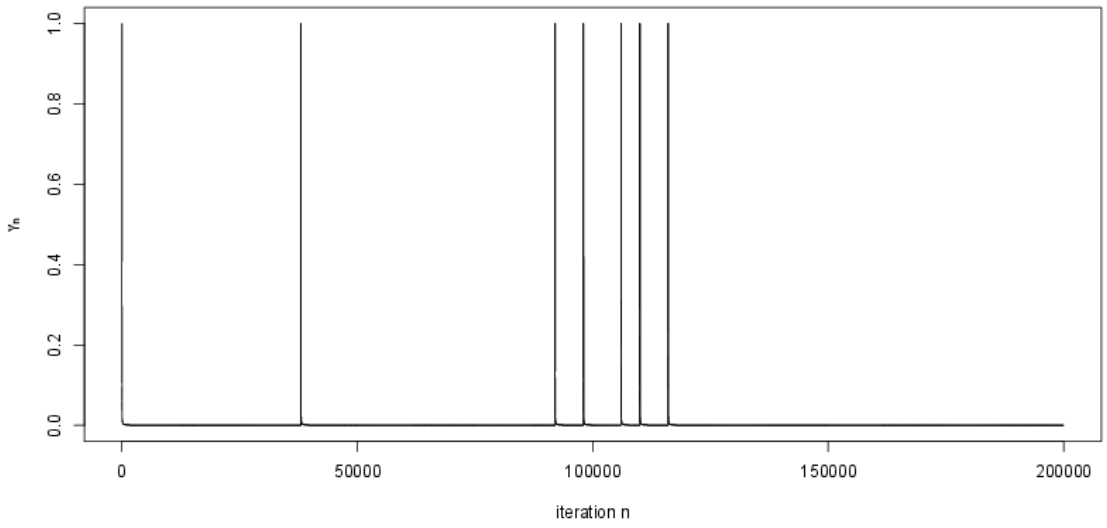


FIGURE 5.7: Step-size sequence for XX using revised AMIS (Algorithm 5.4.1). The sequence is updated around $40,000$ and $90,000$ iterations when the WL chain reaches new regions and the resampling samples from it are used to update the parameters of the AMIS chain.

1. *Add a new component in the mixture proposal distribution and increase the number of proposal components M by 1.*

2. *Set the parameters for the new M-th component as below:*

$$w_M = 1/M, \mu_M = \overline{X^{WL}}, \Sigma_M = Q,$$

where $\overline{X^{WL}}$ and Q can be the sample mean and covariance matrix estimated from WL samples, however, these estimates may be really rough when the number of samples in the new region is small. Hence, an independent sampling algorithm/chain could be used here just for the initialization for the new proposal component.

3. *Update $\{w_j, \mu_j, \Sigma_j\}_{j=1}^{M}$ using the regular formula.*

We apply this second revised algorithm to the previous trimodal target distribution. In Figure 5.8, it can be seen that all modes become well-approximated by components of the proposal mixture distribution, permitting the chain to mix very quickly as indicated by the low autocorrelation and close agreement of the marginal density approximations with the true value.

Still, the disadvantage of this improved algorithm lie in the slow updating of parameters w (Figure 5.9), μ, and Σ. In fact, if we take a closer look at the top panel of plots in Figure 5.8, and compare the contour plot at iteration $150,000$ to that of the target distribution, the center of the contour around $(20, -20)$ are a bit off the true means and are actually not updated much since these proposal components were modified by the sample from WL chain. Therefore, the performance of Algorithm 5.4.2 depends heavily on the initialization obtained from WL chain.

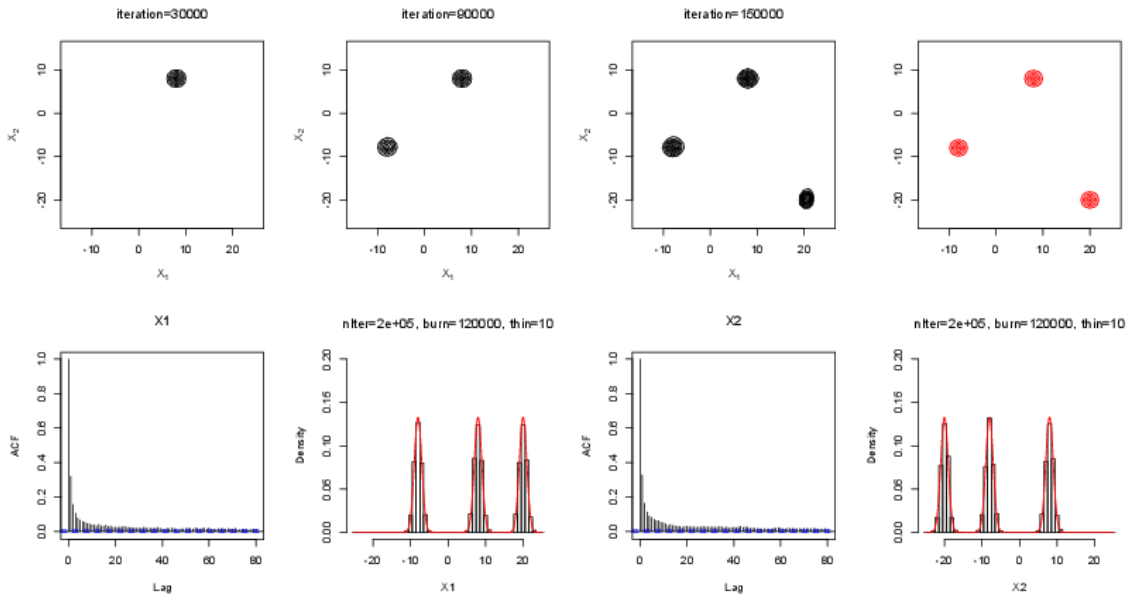


FIGURE 5.8: Simulation results of the XX algorithm using revised AMIS (Algorithm 5.4.2) for the trimodal target distribution. Both the proposal distribution and marginal density approximations agree with the target distribution, and the low autocorrelation imply that the mixing is very good.

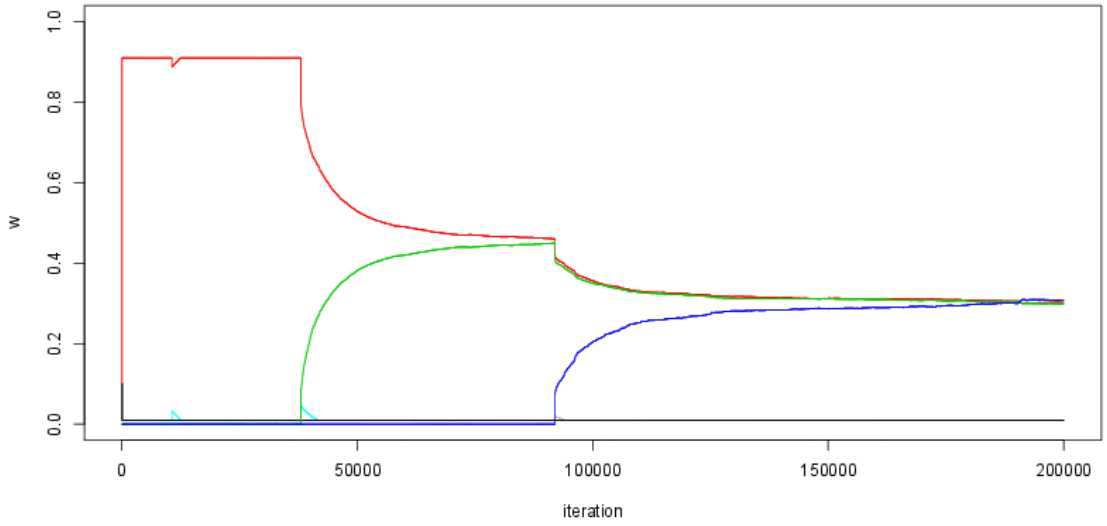


FIGURE 5.9: Weight sequence for XX using revised AMIS (Algorithm 5.4.2). Compared with the result in Figure 5.6 for XX using Algorithm 5.4.1, we can see that the weights for the new dominant components adapt slowly to the true proportions.

5.4.4 *Improvement 3: reset the step-size sequence $\{r_n\}$ and add new proposal components*

From the simulation results in the previous two sections, we see that each of the above two improvements has its own advantage and drawback. It is natural to think about incorporating good aspects from both and make a third improvement.

Algorithm 5.4.3. *When $O_i(X_{n_0}) > D_{th}$:*

1. *Add a new component in the mixture proposal distribution and increase the number of proposal components M by 1.*

2. *Set the parameters for the new M-th component as below:*

$$w_M = 1/M, \mu_M = \overline{X^{WL}}, \Sigma_M = Q,$$

3. *Update $\{w_j, \mu_j, \Sigma_j\}_{j=1}^{M}$ using the regular formula.*

4. *Modify the step-size sequence $\{r_n\}$.*

Once again, we apply Algorithm 5.4.3 to the trimodal target distribution. We see that in Figure 5.10, the mixture proposal distribution approaches to the target distribution really fast, and thus the combination chain generates samples from the target distribution with low autocorrelation. Figure 5.11 describes how quickly the weights can adapt and approximate the true values.

In addition, the user-defined value of M is arbitrary and it is desirable to ensure the algorithm is not handicapped by choosing M too low. The previous modifications (both Algorithm 5.4.2 and Algorithm 5.4.3) of the algorithm allow us to address this easily, by starting with M small and adding one ore more additional components whenever a new mode or region is identified. Note that we need not determine the optimal number of components, which is notoriously difficult, but simply add components as needed to ensure M is sufficiently large.

Up to this point in the chapter, the value of w_{lb} has been fixed as constant. If $w = w_{lb}$, the component is negligible. If $w \in (w_{lb}, 1/M]$, the component is insignificant. And when $w > 1/M$, the component is significant. If we want to increase M automatically, the value assigned to w_{lb} cannot be fixed, and must be updated as well.

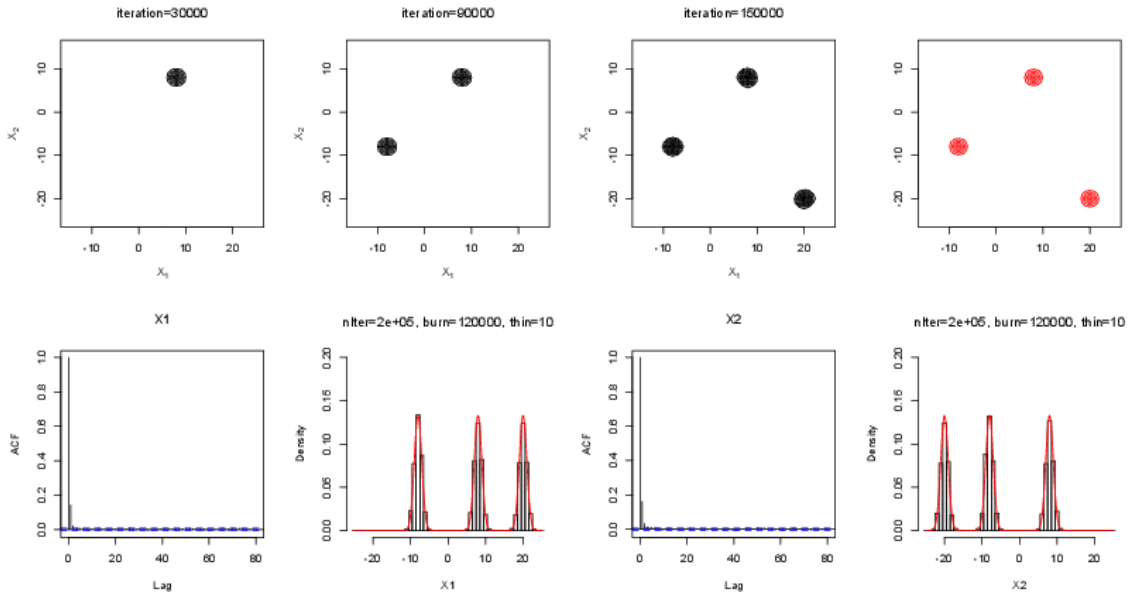


FIGURE 5.10: Simulation results of the XX algorithm using revised AMIS (Algorithm 5.4.2) for the trimodal target distribution. Both the proposal distribution and marginal density approximations agree with the target distribution, and the low autocorrelation imply that the mixing is very good.

5.4.5 Parallel exploration chains

An important aspect of the XX algorithm is the independence of the WL chain dynamics from the AMIS chain. This enables the algorithm to trivially take advantage of multiple processors which are increasingly common in statistical computing environments. Because the exploration process is the most computationally expensive module, and the exploration chains are unaffected by the AMIS component of the XX chain, a natural extension is to run multiple exploration chains in parallel on

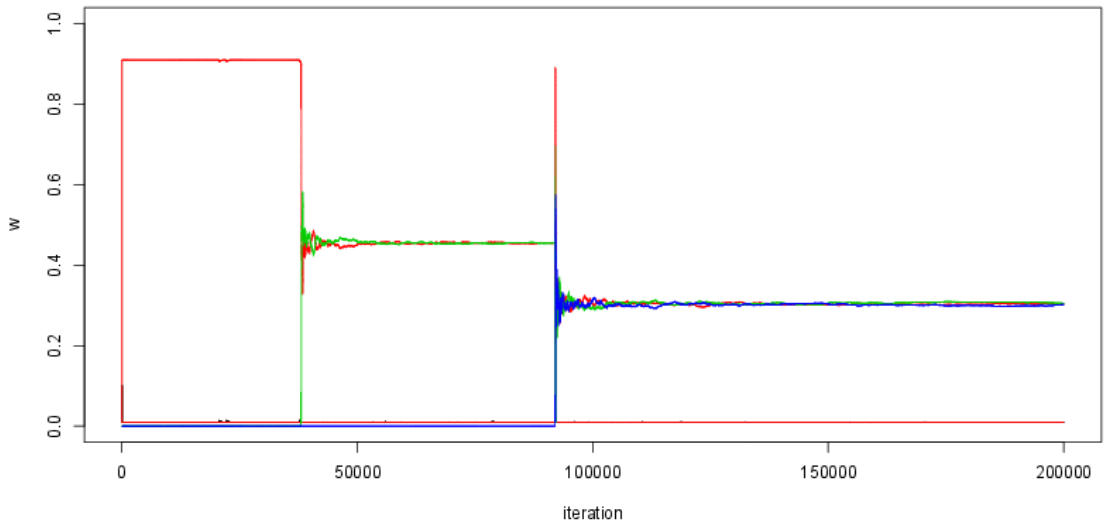


FIGURE 5.11: Weight sequence for XX using revised AMIS (Algorithm 5.4.2). Compared with the result in Figure 5.6 for XX using Algorithm 5.4.1, we can see that the weights for the new dominant components adapt much faster to the true proportions.

separate processors. The associated overhead can be hidden very well because the transfer of samples to updates the AMIS proposal can be performed asynchronously. This approach is used in the second application in Section 5.5.2 below.

5.5 Applications

5.5.1 Mixture exponential regression

Suppose that

$$y_i \sim \begin{cases} \text{Exp}[\theta_1(x_i)] & \text{with probability } \alpha = 0.3, \\ \text{Exp}[\theta_2(x_i)] & \text{with probability } 1-\alpha = 0.7, \end{cases}$$

where $\theta_j(x_i) = \exp(\beta_j^T x_i)$, $j = 1, 2$, $\beta_1 = 1$, $\beta_2 = 6$, and x is all 1's. We would like to make inference on parameters α, β_1, and β_2 based on $\{x_i, y_i\}_{i=1}^n$. This label switching problem serves as a good example to test the performance of the performance of XX

Table 5.2: Parameters for AMIS, WL, and XX algorithms for mixture exponential regression.

Algorithm	Parameters
AMIS	$\lambda = 0.01$, $\tilde{\mu} = (2,2))^T$, $\tilde{\Sigma} = 0.5I$, $M = 10$ $w_{i,1} = 1/M$, $\mu_{i,1} \sim N_2(\mathbf{0}, 2I)$, $\Sigma_{i,1} = 0.25I$, $r_n = 1/n$ $D_{th} = 50$, $r_{\text{restart}} = 1/100$
WL	$d = 10$, $\gamma_n = 1/n$, $\epsilon = 0.3$, $E_{th} = 5$, $n_{\text{split}} = \text{nIter}/10$, $E_{\max} = 1500$
XX	$N_c = \text{nIter}/5, N_{wl} = 100$.

algorithm in a high-dimensional multimodal case. The likelihood function is:

$$L(Y|\alpha,\beta_1,\beta_2) \propto \prod_{i=1}^{n} \left[\frac{\alpha}{\theta_1(x_i)} \exp\left(-\frac{y_i}{\theta_1(x_i)}\right) + \frac{1-\alpha}{\theta_2(x_i)} \exp\left(-\frac{y_i}{\theta_2(x_i)}\right) \right].$$

Assign prior distributions $\pi(\alpha) = \text{Beta}(1,1)$, $\pi(\beta_j) = \text{N}(0,\sigma^2)$, $j = 1,2$, and $\sigma = 10$. Denote the posterior distribution by $\pi(\alpha,\beta_1,\beta_2|Y)$, we can express the energy function as

$$E(\alpha,\beta_1,\beta_2) = -\log(\pi(\alpha,\beta_1,\beta_2|Y)) \propto -\ell(Y|\alpha,\beta_1,\beta_2) + \frac{\beta_1^2+\beta_2^2}{2\sigma^2},$$

where function ℓ is the log likelihood function.

We apply the XX algorithm to sample from this posterior distribution, and compare the results obtained from the AMIS and WL chains, using the parameters in Table 5.2. All three chains were initialized in the same mode at $(\alpha,\beta_1,\beta_2) = (0.3,1,6)$. We see in Figure 5.12 that the AMIS algorithm by itself fails to escape from the initial mode. The WL chain in Figure 5.13 is able to cross over to the second mode, but the autocorrelation is very strong and the resulting density approximations are somewhat rough. Figure 5.14 shows the significantly improved convergence and rapid decay of autocorrelation for the XX algorithm (using revised AMIS Algorithm 5.4.2), with resulting improvement in the density approximation. Figure 5.15 shows how the weight sequences are quickly updated.

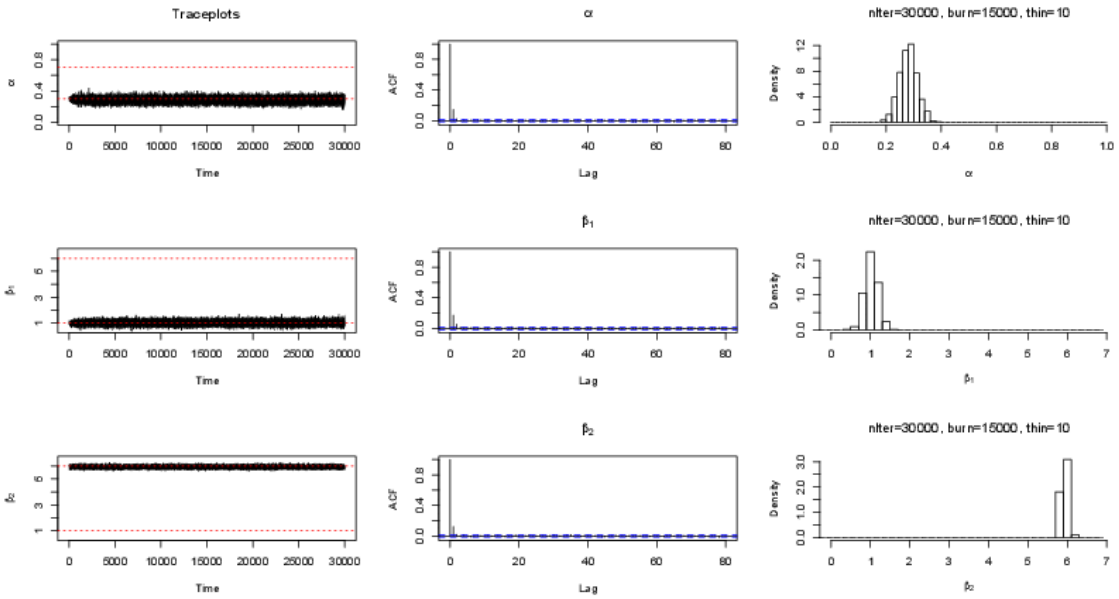


FIGURE 5.12: Simulation results of the AMIS algorithm for the mixture exponential regression problem. The AMIS sampler failed to escape from the initial local mode. Therefore, the low autocorrelations are misleading.

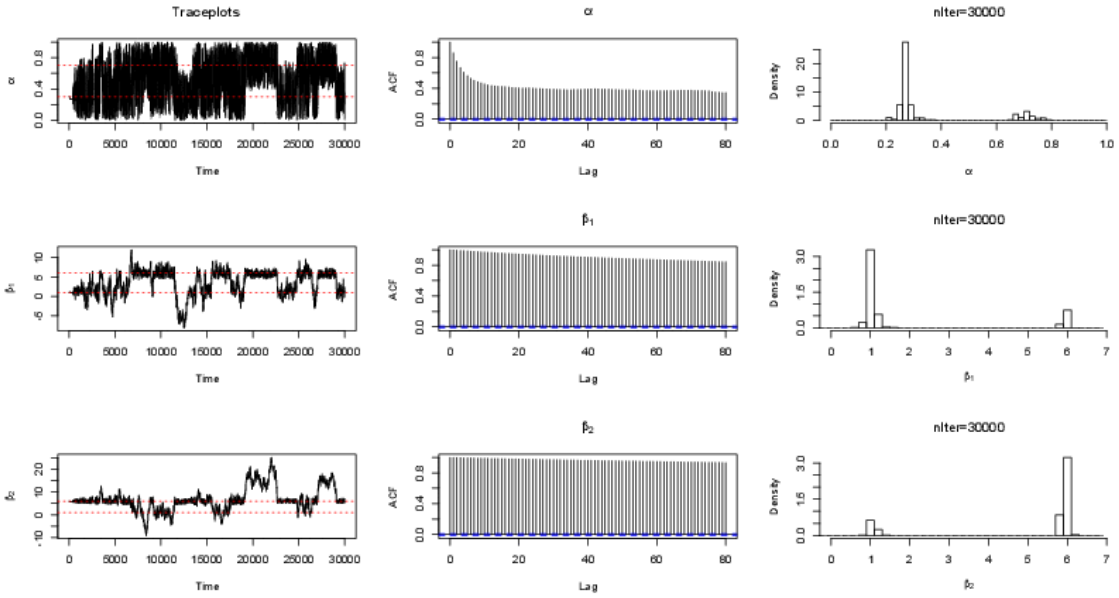


FIGURE 5.13: Simulation results of the WL algorithm for the mixture exponential regression problem. The WL chain can cross the energy barrier to reach the other mode, but samples from this chain are highly correlated, and we can only use the resampling samples to approximate the target distributions.

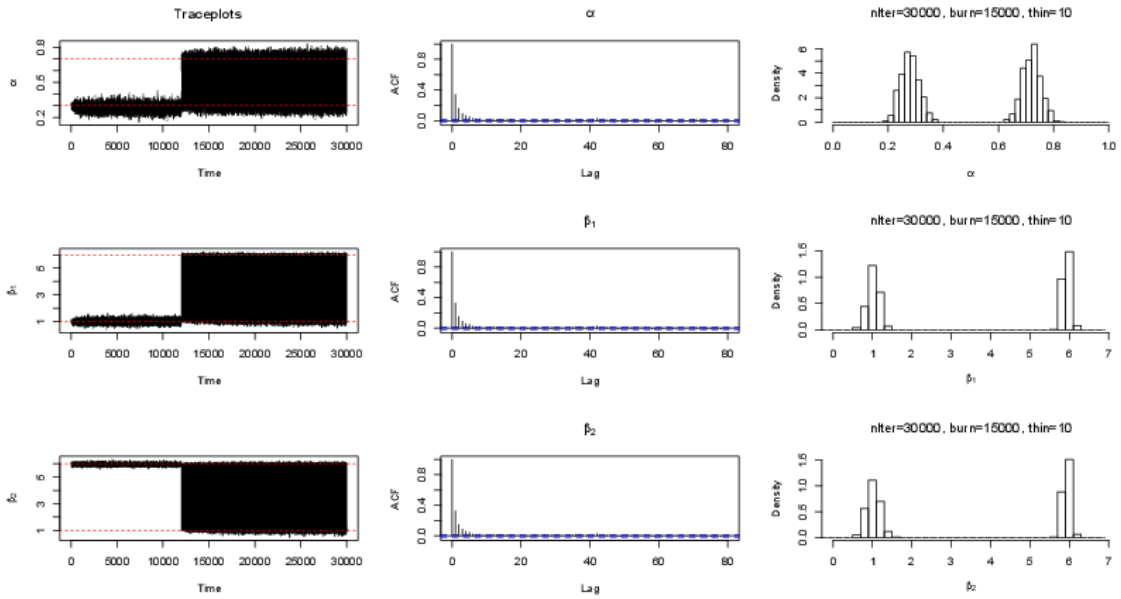


FIGURE 5.14: Simulation results of the XX algorithm (using revised AMIS Algorithm 5.4.3) for the mixture exponential regression problem. After borrowing information from the WL chain around iteration 10,000, the mixing of the sampler is very good, and the posterior modes of the parameters agree with the true values in simulation.

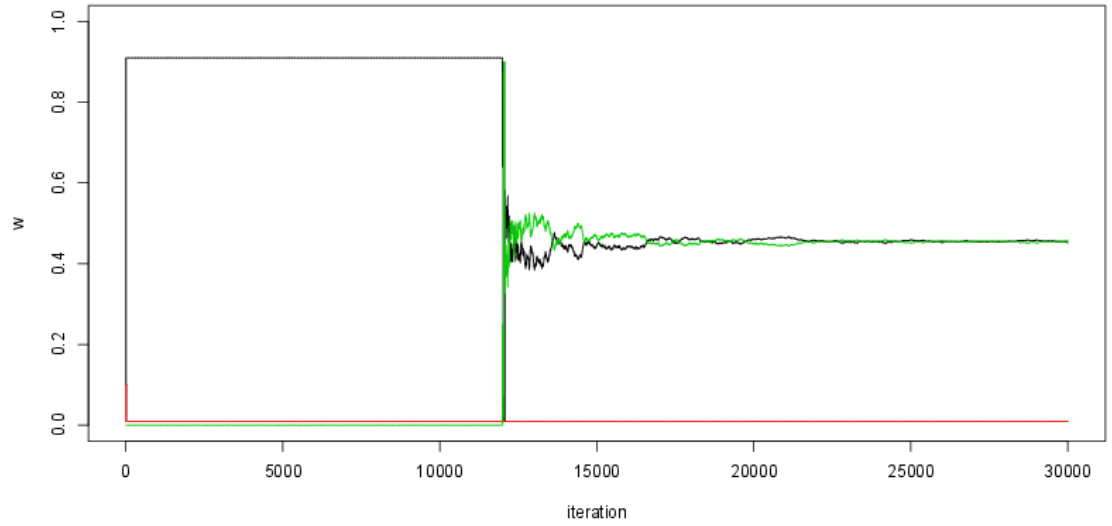


FIGURE 5.15: Weights sequence of the XX chain (using revised AMIS Algorithm 5.4.3) for the mixture exponential regression problem.

5.5.2 *Bayesian neural network analysis*

Nonlinear regression and classification using neural network models is a widely used tool in machine learning and statistical pattern recognition (Ripley, 1996; Hastie et al., 2001). Bayesian approaches have notable advantages in the use of priors to control model complexity and marginalization over posterior uncertainty to improve predictive performance (Neal, 1997; Lee, 1998; MacKay, 1992, 1995). However, neural network fitting by optimization is notoriously subject to multiple minima, and this translates to highly multimodal posterior distributions in a Bayesian context. Sampling from such posteriors is highly non-trivial and significant research in MCMC methods has been aimed at this problem (Neal, 1997). Here we show the advantages of the XX framework on such problems.

Let $(\mathbf{y}_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, be the observed data from the model

$$y_{ij} = \beta_{j0} + \sum_{h=1}^{H} \beta_{jh} \Phi_{jh}(\gamma_{jh0} + \mathbf{x}_i^T \gamma_{jh}) + \epsilon_{ij},$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$, $\mathbf{x}_i \in \mathbb{R}^K$ is the input variable, $\mathbf{y}_i \in \mathbb{R}^J$ is the multivariate response, and $\Phi(t) = 1/(1 + e^{-t})$ is the logistic function. We simulated a dataset of size $n = 20$ from this model with $J = 1$, $H = 2$, and $K = 2$. The parameters are

$$\theta = (\beta_{10}, \beta_{11}, \beta_{12}, \gamma_{110}, \gamma_{120}, \boldsymbol{\gamma}_{11} = (\gamma_{111}, \gamma_{112}), \boldsymbol{\gamma}_{12} = (\gamma_{121}, \gamma_{122}), \sigma).$$

Input variables were sampled independently from $x_{i1} \sim \text{Unif}((-1.932, -0.453) \cup (0.453, 1.932))$ and $x_{i2} \sim \text{Unif}(0.534, 3.142)$

Table 5.3: Simulation parameters for the neural network problem.

θ	β_{10}	β_{11}	β_{12}	γ_{110}	γ_{120}	$\boldsymbol{\gamma}_{11}$	$\boldsymbol{\gamma}_{12}$	σ
True	1	2	1	-3	2	(1,2)	(-1,1)	0.05

Notice that the logistic function $\Phi(t)$ is pretty flat when $|t| > 5$. Hence, for fixed parameters γ_{1h0} and γ_{1h}, if different input values $\mathbf{x}_i$ and $\mathbf{x}_{i'}$ make $t_i = \gamma_{1h0} + \mathbf{x}_i^T \gamma_{1h}$ and

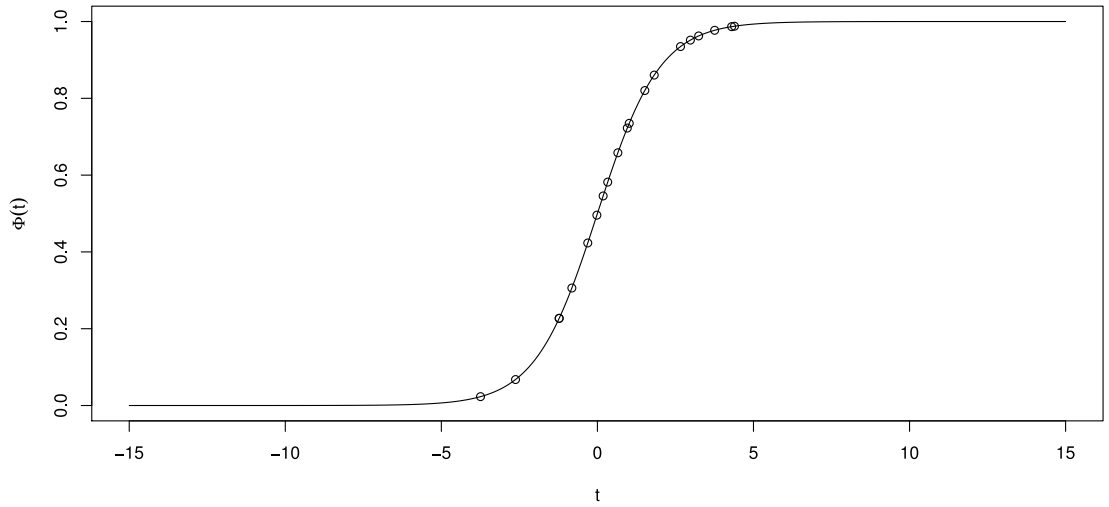
$t_{i'} = \gamma_{1h0} + \mathbf{x}_{i'}^T \gamma_{1h}$ both greater than 5 or less than -5, then $\Phi(t_i)$ and $\Phi(t_{i'})$ are close to each other and $\mathbf{y}_i$ is similar to $\mathbf{y}_{i'}$ as a result. When different $\mathbf{x}$ values correspond to similar $\mathbf{y}$ values, there might not be enough information to make inference for the parameter. Consequently, there might be too many modes in the parameter space. In the simulation, we set the parameters of θ according to Table 5.3, which are selected to make $\Phi(\gamma_{1h0} + \mathbf{x}_i^T \gamma_{1h})$ distinct for different $\mathbf{x}_i$'s. Particularly, the values for $h = 1$ and $h = 2$ are drawn in Figures 5.16a and 5.16b, respectively. The resulting dataset is shown in Figure 5.17.

There are several sources of multimodality in this example. First, the likelihood function is fairly flat. That is, many values of the likelihood at different parameters are nearly equal, indicating this likelihood may have many maximum points. Another source is an obvious "label switching" problem familiar from mixture modeling: switching the h parameters of the β's and γ's yields the same likelihood. A third one comes from the anti-symmetry ($\Phi(t) = 1 - \Phi(-t)$) of the logistic function such that

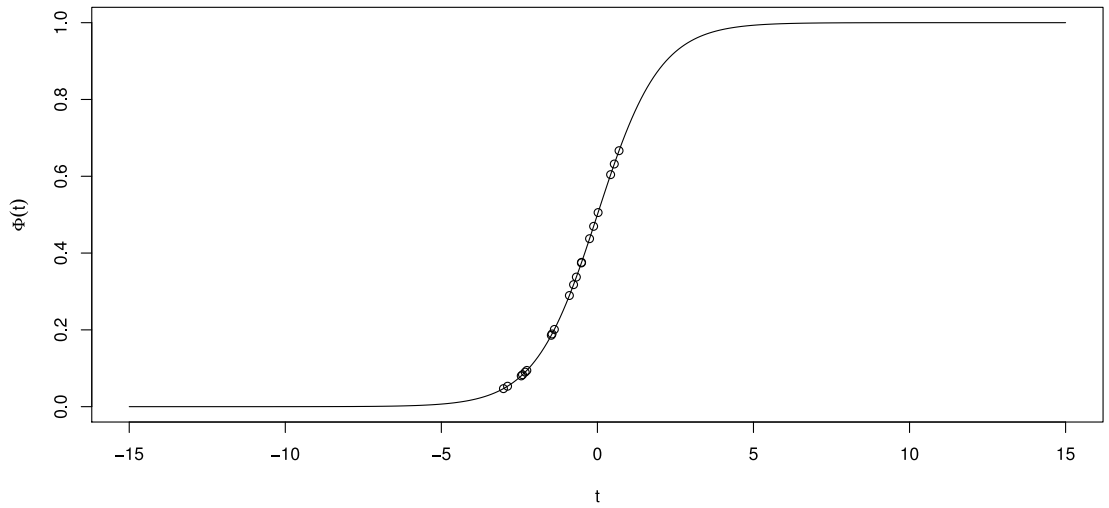
$$\beta_{j0} + \beta_{j1}\Phi(t) = (\beta_{j0} + \beta_{j1}) - \beta_{j1}\Phi(-t).$$

For instance, the likelihood for parameter $\theta = (\beta_{10}, \beta_{11}, \beta_{12}, \gamma_{110}, \gamma_{120}, \boldsymbol{\gamma}_{11}, \boldsymbol{\gamma}_{12}, \sigma)$ is the same as that for parameter $\theta' = (\beta_{10} + \beta_{11}, -\beta_{11}, \beta_{12}, -\gamma_{110}, \gamma_{120}, -\boldsymbol{\gamma}_{11}, \boldsymbol{\gamma}_{12}, \sigma)$. As a result, we expect at lest eight significant modes in the model by symmetry, making it an excellent test bed for the XX algorithm when n is sufficiently large to make crossing between the modes challenging. Table 5.4 shows the locations of these modes under the true parameter, which should be approximately the locations of the posterior modes (due to finite value of n).

We assign prior distributions to the parameters as follows: $\beta_{\cdot\cdot} \overset{\text{iid}}{\sim} N(0, 10^2)$, $\gamma_{\cdot\cdot} \overset{\text{iid}}{\sim} N(0, 10^2)$, and $\sigma^2 \sim \text{IG}(2, 0.5)$. As before, we compare the XX algorithm against the two component algorithms applied individually. The parameters are



(a) $\Phi(-3 + x_{i1} + 2x_{i2}), \quad i = 1, \ldots, 20.$



(b) $\Phi(2 - x_{i1} + x_{i2}), \quad i = 1, \ldots, 20.$

FIGURE 5.16: $\Phi(\gamma_{1h0} + \mathbf{x}_i^T \gamma_{1h})$ values for $h = 1$ and $h = 2$ respctively.

given in Table 5.5. Each algorithm was run for $3e5$ iterations, initialized at a point uniformly drawn from the state space. In this high dimensional example, it takes a long time for the WL chain to explore the entire state space, and we use the parallel exploration XX algorithm described in Section 5.4.5, using 15 independent WL chains in parallel for the exploration component.

For the AMIS algorithm, the mixture proposal distribution shows only one signif-

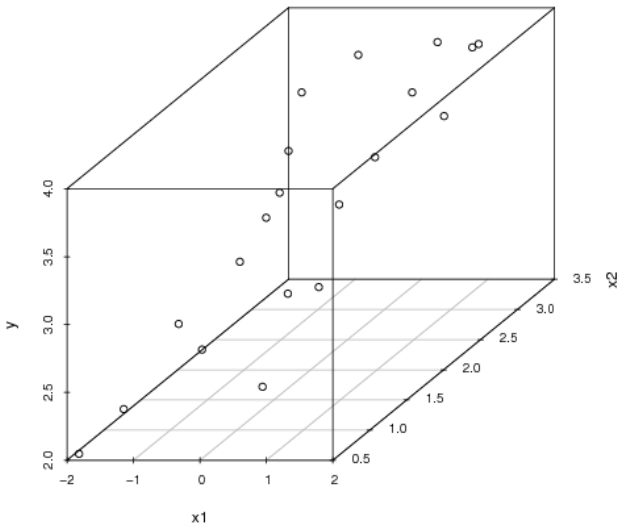


FIGURE 5.17: Simulated dataset for the neural network example.

Table 5.4: One set of the true parameters used in simulating data and seven other sets of parameters yielding the same likelihood. These points are approximately the locations of the modes.

θ	β_{10}	β_{11}	β_{12}	γ_{110}	γ_{120}	$\boldsymbol{\gamma}_{11}$	$\boldsymbol{\gamma}_{12}$	σ	index
True	1	2	1	-3	2	(1, 2)	(-1, 1)	0.05	1
	1	1	2	2	-3	(-1, 1)	(1, 2)	0.05	2
	3	-2	1	3	2	(-1, -2)	(-1, 1)	0.05	3
	3	1	-2	2	3	(-1, 1)	(-1, -2)	0.05	4
	2	2	-1	-3	-2	(1, 2)	(1, -1)	0.05	5
	2	-1	2	-2	-3	(1, -1)	(1, 2)	0.05	6
	4	-2	-1	3	-2	(-1, -2)	(1, -1)	0.05	7
	4	-1	-2	-2	3	(1, -1)	(-1, -2)	0.05	8

Table 5.5: Parameters for AMIS, WL, and XX algorithms for the neural network problem.

Algorithm	Parameters
AMIS	$\lambda = 0.01$, $\tilde{\mu} = (2,2))^T$, $\tilde{\Sigma} = 0.5I$, $M = 30$ $w_{i,1} = 1/M$, $\mu_{i,1} \sim N_2(\mathbf{0}, 2I)$, $\Sigma_{i,1} = 0.25I$, $r_n = 1/n$ $D_{th} = 50$, $r_{\text{restart}} = 1/100$, $w_{lb} = 0.001$
WL	$d = 15$, $\gamma_n = 1/n$, $\epsilon = 0.3$, $E_{th} = 5$, $n_{\text{split}} = \text{nIter}/10$, $E_{\max} = 3000$
XX	$N_c = \text{nIter}/10$, $N_{wl} = 100$, $D_{th} = 1/(2 \cdot w_{lb})$, adaptive r_{restart}

icant component throughout the entire run, which depends on the starting point and $\mu_{i,1}$. The parallel WL chains each discover a different set of modes within the simulation length depending on its initial point (see Table 5.6). In contrast, the mixture proposal distribution for the XX algorithm adapts to capture all eight significant modes as desired. Marginal posterior distributions of the parameters are shown in Figure 5.18, with the (approximate) modes of Table 5.4 (known by symmetry arguments) shown in red dashed lines. We see excellent coverage by the sampling. Bivariate posterior distribution for parameter pairs $(\beta_{11}, \gamma_{120})$ and $(\gamma_{111}, \gamma_{121})$ are shown in Figure 5.19, with the expected 8 and 4 modes clearly visible, respectively.

5.6 Discussion

In this chapter, we propose an exploration/exploitation algorithm, referred to as XX algorithm, to address the convergence and autocorrelation issues in the adaptive MCMC. The exploration component of the XX algorithm is based on the AMIS algorithm while the exploitation component is based on the WL algorithm introduced in Chapter 4. We show that for cases where modes are far apart, combining AMIS and WL in the framework of XX exhibits clear advantage over either component applied individually. We also present one example in which combining the original AMIS with WL is superior to the standalone AMIS, but inferior to WL, caused by a late escape across the energy boundary of the WL chain. We subsequently suggest several

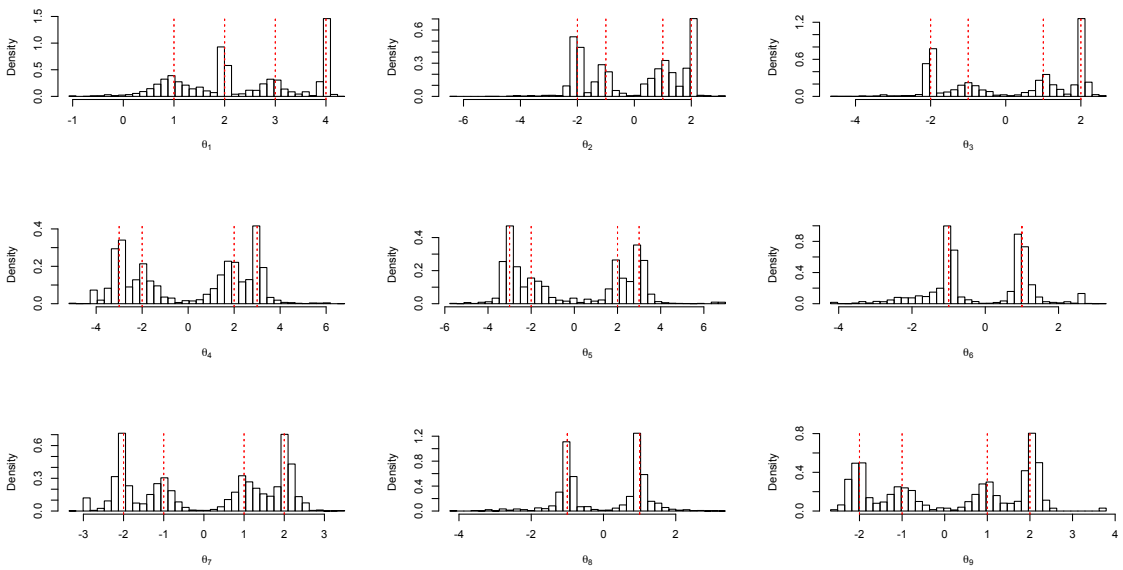


FIGURE 5.18: Marginal distributions of the samples from XX chain for the neural network example. In each plot, the local modes agree with the numbers shown in Table 5.4.

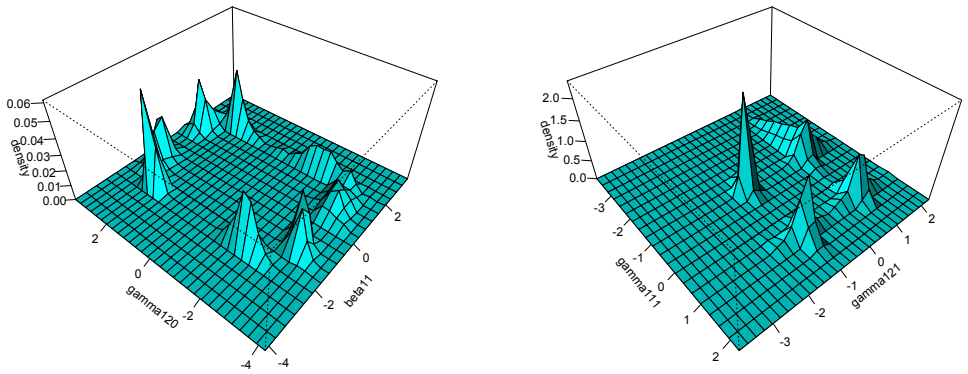


FIGURE 5.19: Bivariate distributions of $(\beta_{11}, \gamma_{120})$, and $(\gamma_{111}, \gamma_{121})$. We can compare the locations of the modes with the values shown in Table 5.4.

Table 5.6: Indices and orders of the modes visited by the 15 WL chains, respectively. See Table 5.4 for the values in the second and third columns. Each chain found a different set of modes.

WL chain	First mode found	Other modes found later
1	5	8
2	7	
3	4	8,3
4	8	4,7
5	8	4
6	2	6,1
7	5	6
8	8	3
9	5	6
10	5	2,1
11	6	7,5
12	5	
13	3	7,8
14	8	7
15	7	8,3

improvements and demonstrate their performance with several examples. Remaining work on this topic includes testing the method in much higher dimensions.

Appendix A

Simulation Method for Non-Stationary Negative Binomial Branching Process

1. The Data Augmentation method introduced in Section **??** is also a simulation method for the stationary Negative Binomial Branching process, bNB(α, β, ρ). However, this simulation method does not work for $X_t \sim \text{bNB}(\alpha_t, \beta, \rho)$ by just modifying the subscripts as below, as the marginal distributions of $\{X_t\}$ are not NB(α_t, β) and the correlation is not ρ.

$$X_0 \sim \text{NB}\left(\alpha_0, \frac{\beta}{1+\beta}\right)$$

$$Y_t \sim \text{Bi}\left(x_{t-1}, \frac{\rho\beta}{1+\beta-\rho}\right), \quad \zeta_t \sim \text{NB}\left(\alpha_t + y_t, \frac{\beta}{1+\beta-\rho}\right)$$

$$X_t = y_t + \zeta_t, \qquad 1 \leqslant t \leqslant T.$$

2. First, we propose a new strategy to simulate non-stationary negative binomial processes with varying parameter α_t, such that the marginal distributions of $\{X_t\}$ are NB(α_t, β):

$$X_0 \sim \text{NB}\left(\alpha_0, \frac{\beta}{1+\beta}\right)$$

(1) If $\alpha_t = \alpha_{t-1}$:

$$Y_t \sim \text{Bi}\left(x_{t-1}, \frac{\rho\beta}{1+\beta-\rho}\right), \quad \zeta_t \sim \text{NB}\left(\alpha_t + y_t, \frac{\beta}{1+\beta-\rho}\right), \quad X_t = y_t + \zeta_t$$

(2) If $\alpha_t > \alpha_{t-1}$:

$$Y_t \sim \text{Bi}\left(x_{t-1}, \frac{\rho\beta}{1+\beta-\rho}\right), \quad \zeta_t \sim \text{NB}\left(\alpha_{t-1} + y_t, \frac{\beta}{1+\beta-\rho}\right)$$

$$X_t = y_t + \zeta_t + X^*, \quad \text{where } X^* \perp X_t \text{ and } X^* \sim \text{NB}\left(\alpha_t - \alpha_{t-1}, \frac{\beta}{1+\beta}\right)$$

(3) If $\alpha_t < \alpha_{t-1}$:

(*a*) $p^* \sim \text{Be}(\alpha_t, \alpha_{t-1} - \alpha_t), \quad X^*_{t-1} \sim \text{Bi}(x_{t-1}, p^*)$

(*b*) $Y_t \sim \text{Bi}\left(x^*_{t-1}, \frac{\rho\beta}{1+\beta-\rho}\right), \quad \zeta_t \sim \text{NB}\left(\alpha_t + y_t, \frac{\beta}{1+\beta-\rho}\right), \quad X_t = y_t + \zeta_t$

3. Proof:

   (1) If $\alpha_t = \alpha_{t-1}$, use the same approach as for stationary Negative Binomial branching process.

   (2) If $\alpha_t > \alpha_{t-1}$, it follows from the property of negative binomial distribution: "The sum of independent negative-binomially distributed random variables $\text{NB}(r_1, p)$ and $\text{NB}(r_2, p)$ is negative-binomially distributed with the same p but with "r-value" $r_1 + r_2$".

   (3) If $\alpha_t < \alpha_{t-1}$, see proof in Appendix B.

4. Calculation of $\text{Cor}(X_{t-1}, X_t)$ for $1 \leqslant t \leqslant T$:

   (a) If $\alpha_t = \alpha_{t-1}$, $\text{Cor}(X_{t-1}, X_t) = \rho$ same as in the stationary process.

(b) If $\alpha_t > \alpha_{t-1}$, according to the new strategy of simulation:

$$X_{t-1} \sim \mathrm{NB}\left(\alpha_{t-1}, \tfrac{\beta}{1+\beta}\right), \quad \mu_{t-1} = \tfrac{\alpha_{t-1}}{\beta}, \quad \sigma_{t-1} = \tfrac{\sqrt{\alpha_{t-1}(1+\beta)}}{\beta},$$
$$X_t \sim \mathrm{NB}\left(\alpha_t, \tfrac{\beta}{1+\beta}\right), \qquad \mu_t = \tfrac{\alpha_t}{\beta}, \qquad \sigma_t = \tfrac{\sqrt{\alpha_t(1+\beta)}}{\beta}.$$

Denote $V_t = Y_t + \zeta_t$, then

$$V_t \sim \mathrm{NB}\left(\alpha_{t-1}, \frac{\beta}{1+\beta}\right), \quad \mathrm{Cor}(X_{t-1}, V_t) = \rho, \quad X_{t-1} \perp X^*.$$

$$\begin{aligned}
&\mathrm{E}[X_{t-1}X_t] \\
=&\mathrm{E}[X_{t-1}(V_t + X^*)] \\
=&\mathrm{E}[X_{t-1}V_t] + \mu_{t-1} \cdot \frac{\alpha_t - \alpha_{t-1}}{\beta} \\
=&\rho\sigma_{t-1}^2 + \mu_{t-1}^2 + \mu_{t-1}(\mu_t - \mu_{t-1}) \\
=&\rho\sigma_{t-1}^2 + \mu_{t-1}\mu_t.
\end{aligned}$$

So $\mathrm{Cor}(X_{t-1}, X_t) = \dfrac{\mathrm{E}[X_{t-1}X_t] - \mu_{t-1}\mu_t}{\sigma_{t-1}\sigma_t} = \rho\dfrac{\sigma_{t-1}}{\sigma_t} = \boldsymbol{\rho}\sqrt{\dfrac{\boldsymbol{\alpha_{t-1}}}{\boldsymbol{\alpha_t}}} < \rho.$

(3) If $\alpha_t < \alpha_{t-1}$, we can look at this as a 'reverse step' of (2) to get

$$\mathrm{Cor}(X_{t-1}, X_t) = \boldsymbol{\rho}\sqrt{\frac{\boldsymbol{\alpha_{t-1}}}{\boldsymbol{\alpha_t}}} < \rho.$$

Alternatively, according to the new strategy of simulation and Appendix B, we have:

$$\begin{aligned}
X_{t-1} &\sim \mathrm{NB}\left(\alpha_{t-1}, \tfrac{\beta}{1+\beta}\right), & X^*_{t-1} &\sim \mathrm{NB}\left(\alpha_t, \tfrac{\beta}{1+\beta}\right), \\
X^{**}_{t-1} &\sim \mathrm{NB}\left(\alpha_{t-1} - \alpha_t, \tfrac{\beta}{1+\beta}\right), & X_t &\sim \mathrm{NB}\left(\alpha_t, \tfrac{\beta}{1+\beta}\right) \\
X_{t-1} &= X^*_{t-1} + X^{**}_{t-1}, & X^{**}_{t-1} &\perp X^*_{t-1}, \\
X^{**}_{t-1} &\perp X_t, & \mathrm{Cor}(X^*_{t-1}, X_t) &= \rho.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\mathrm{E}[X_{t-1}X_t] \\
=&\mathrm{E}[X_t(X^*_{t-1} + X^{**}_{t-1})] \\
=&\mathrm{E}[X_t X^*_{t-1}] + \mu_t \cdot \frac{\alpha_{t-1} - \alpha_t}{\beta} \\
=&\rho\sigma_t^2 + \mu_t^2 + \mu_t(\mu_{t-1} - \mu_t) \\
=&\rho\sigma_t^2 + \mu_{t-1}\mu_t.
\end{aligned}$$

$$\text{So } \mathrm{Cor}(X_{t-1}, X_t) = \frac{\mathrm{E}[X_{t-1}X_t] - \mu_{t-1}\mu_t}{\sigma_{t-1}\sigma_t} = \rho\frac{\sigma_t}{\sigma_{t-1}} = \boldsymbol{\rho}\sqrt{\frac{\boldsymbol{\alpha_t}}{\boldsymbol{\alpha_{t-1}}}} < \rho.$$

5. Then, we modify the new strategy to keep ρ constant:

   For $1 \leqslant t \leqslant T$, let $X_0 \sim \mathrm{NB}\left(\alpha_0, \dfrac{\beta}{1+\beta}\right)$ and

   (1) If $\alpha_t = \alpha_{t-1}$:

   $$Y_t \sim \mathrm{Bi}\left(x_{t-1}, \frac{\rho\beta}{1+\beta-\rho}\right), \quad \zeta_t \sim \mathrm{NB}\left(\alpha_t + y_t, \frac{\beta}{1+\beta-\rho}\right), \quad X_t = y_t + \zeta_t;$$

   (2) If $\alpha_t > \alpha_{t-1}$:

   $$Y_t \sim \mathrm{Bi}\left(x_{t-1}, \frac{\rho'\beta}{1+\beta-\rho'}\right), \quad \zeta_t \sim \mathrm{NB}\left(\alpha_{t-1} + y_t, \frac{\beta}{1+\beta-\rho'}\right),$$

   where $\boldsymbol{\rho'} = \boldsymbol{\rho}\sqrt{\dfrac{\boldsymbol{\alpha_t}}{\boldsymbol{\alpha_{t-1}}}} > \rho$, and

   $$X_t = y_t + \zeta_t + X^*, \quad \text{where } X^* \sim \mathrm{NB}\left(\alpha_t - \alpha_{t-1}, \frac{\beta}{1+\beta}\right);$$

   (3) If $\alpha_t < \alpha_{t-1}$:

   $$(a)\ p^* \sim \ \mathrm{Be}(\alpha_t, \alpha_{t-1} - \alpha_t), \quad X^*_{t-1} \sim \ \mathrm{Bi}(x_{t-1}, p^*)$$

(*b*) $Y_t \sim \text{Bi}\left(x^*_{t-1}, \dfrac{\rho'\beta}{1+\beta-\rho'}\right), \quad \zeta_t \sim \text{NB}\left(\alpha_t + y_t, \dfrac{\beta}{1+\beta-\rho'}\right),$

where $\boldsymbol{\rho'} = \boldsymbol{\rho}\sqrt{\dfrac{\boldsymbol{\alpha_{t-1}}}{\boldsymbol{\alpha_t}}} > \rho$, and $X_t = y_t + \zeta_t$.

(4) To make sure $\rho' \leqslant 1$, $\{\alpha_t\}$ must satisfy $1 < \dfrac{\alpha_t}{\alpha_{t-1}} \leqslant \dfrac{1}{\rho^2}$ or $\rho^2 \leqslant \dfrac{\alpha_t}{\alpha_{t-1}} < 1$.

Appendix B

Beta-Binomial Distribution

1. The Beta-Binomial distribution is a discrete probability distribution on a finite support of non-negative integers. It is frequently used when the probability of success in each of a fixed or known number of Bernoulli trials is either unknown or random.

   If a random variable is drawn from the Beta-Binomial distribution, namely $X \sim BB(n, \alpha, \beta)$, then the probability density function is:

$$f(X = x \mid n, \alpha, \beta) = \left(\begin{array}{c} n \\ x \end{array} \right) \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha + x)\Gamma(\beta + n - x)}{\Gamma(\alpha + \beta + n)}, \quad 0 \leqslant x \leqslant n.$$

2. Beta-Binomial Distribution can be considered as a compound distribution. To sample from $\mathrm{BB}(n, \alpha, \beta)$, we can proceed in the following two steps:

$$p \sim Be(\alpha, \beta)$$

$$x \sim Bi(n, p).$$

Proof:

$$\begin{aligned}\Pr[X=x] &= \int_0^1 \mathrm{Bin}(x \mid n,p)\cdot \mathrm{Beta}(p \mid \alpha,\beta)\\ &= \int_0^1 \binom{n}{x} p^x q^{n-x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}q^{\beta-1} d\,p, \quad q=1-p\\ &= \binom{n}{x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+x)\Gamma(\beta+n-x)}{\Gamma(\alpha+\beta+n)}, \quad 0\leqslant x\leqslant n.\end{aligned}$$

3. In Appendix A, we want to generate a random variable X_{t-1}^*, given $X_{t-1} \sim \mathrm{NB}(\alpha_{t-1},p)$, such that $X_{t-1}=X_{t-1}^*+X_{t-1}^{**}$ and

$$X_{t-1}^* \sim \mathrm{NB}(\alpha_{t-1},p) \perp X_{t-1}^{**} \sim \mathrm{NB}(\alpha_{t-1}-\alpha_t,p), \quad \alpha_t<\alpha_{t-1}.$$

The conditional pmf of $x=X_{t-1}^*$, given $y=X_{t-1}$, is:

$$\begin{aligned}f(x|y)=\frac{f(x,y)}{f(y)} &= \frac{\frac{\Gamma(\alpha_t+x)}{\Gamma(\alpha_t)x!}p^{\alpha_t}q^x \cdot \frac{\Gamma(\alpha_{t-1}-\alpha_t+y-x)}{\Gamma(\alpha_{t-1}-\alpha_t)(y-x)!}p^{\alpha_{t-1}-\alpha_t}q^{y-x}}{\frac{\Gamma(\alpha_{t-1}+y)}{\Gamma(\alpha_{t-1})y!}p^{\alpha_{t-1}}q^y}\\ &= \binom{y}{x}\frac{\Gamma(\alpha_{t-1})}{\Gamma(\alpha_t)\Gamma(\alpha_{t-1}-\alpha_t)}\frac{\Gamma(\alpha_t+x)\Gamma(\alpha_{t-1}-\alpha_t+y-x)}{\Gamma(\alpha_{t-1}+y)}\\ &\sim \mathrm{BB}(n=y,\alpha=\alpha_t,\beta=\alpha_{t-1}-\alpha_t).\end{aligned}$$

Hence, sample

$$p^* \sim Be(\alpha_t,\alpha_{t-1}-\alpha_t), \quad X_{t-1}^* \sim Bi(x_{t-1},p^*).$$

Appendix C

Posterior Distribution

1. Improper prior distributions for μ and uniform prior for ρ:

$$\pi(\mu) \propto \frac{1}{\mu}, \quad 0 < \mu < \infty; \qquad \pi(\rho) = 1, \quad 0 \leqslant \rho < 1.$$

2. Likelihood function:

$$\begin{aligned}
&\mathrm{P}[\boldsymbol{Y} = y \mid \alpha, \rho, p] = p(y_1) \prod_{1<j\leqslant n} p(y_j \mid y_{j-1}) \\
=&\frac{\Gamma(\alpha + y_1)}{\Gamma(\alpha) y_1!} p^{\alpha} (1-p)^{y_1} \cdot \prod_{1<j\leqslant n} \frac{\Gamma(\alpha + y_j)}{\Gamma(\alpha) y_j!} \cdot \frac{p^{\alpha}(1-\rho)^{y_j + y_{j-1}}(1-p)^{y_j}}{(1-\rho+\rho p)^{\alpha + y_j + y_{j-1}}} \\
&\cdot \sum_{\xi=0}^{y_j \wedge y_{j-1}} \frac{y_{j-1}!\Gamma(\alpha) y_j!}{(y_{j-1}-\xi)!\Gamma(\alpha+\xi)(y_j-\xi)!\xi!} \left(\frac{\rho}{(1-\rho)^2} \frac{p^2}{(1-p)} \right)^{\xi},
\end{aligned}$$

where $\mu = \dfrac{\alpha}{\beta}$, $p = \dfrac{\beta}{1+\beta}$.

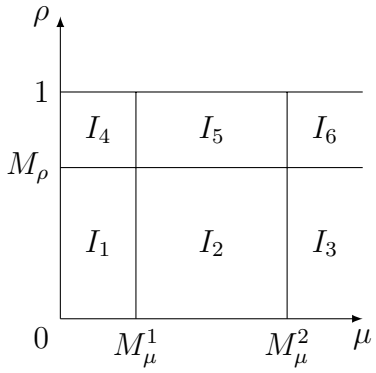
3. We want to show that the posteriors distribution

$$\pi(\mu, \rho \mid \boldsymbol{Y} = y) = \pi(\mu)\pi(\rho)\mathrm{P}[\boldsymbol{Y} = y \mid \alpha, \rho, p]$$

is proper:

$$\int_0^1 \int_0^\infty \pi(\mu, \rho \mid \boldsymbol{Y} = y)\, d\mu d\rho < \infty$$

4. Divide the $\mu\rho$-plane up into 9 rectangles and the integral into 6 double integrals over the corresponding sets:



where M_μ^1 is a small value close to 0, M_μ^2 is a large value and M_ρ is a value close to 1.

5.

$$I_2 = \int_0^{M_\rho} \int_{M_\mu^1}^{M_\mu^2} \pi(\mu, \rho \mid \boldsymbol{Y} = y)\, d\mu d\rho.$$

$\pi(\mu, \rho \mid \boldsymbol{Y} = y)$ is a continuous function in μ and ρ over the bounded set $[0, M_\rho] \times [M_\mu^1, M_\mu^2]$, so the posterior is integrable.

6.

$$I_1 = \int_0^{M_\rho} \int_0^{M_\mu^1} \pi(\mu, \rho \mid \boldsymbol{Y} = y)\, d\mu d\rho$$

On the set $[0, M_\rho] \times (0, M_\mu^1]$, both ρ and $\dfrac{\rho}{(1-\rho)^2}$ are bounded. The range of rockfall data Y is $[0, 212]$, so all terms only involving $\{y_i\}_{i=1}^n$ are all bounded.

Moreover, the exponents of p, $1-p$, $1-\rho$ and $\dfrac{p}{1-\rho+\rho p}$, which are all less than 1, are also bounded by 1. Hence,

$$
\begin{aligned}
&\pi(\mu,\rho \mid \boldsymbol{Y}=y)\\
\propto&\frac{1}{\alpha}\mathrm{P}[\boldsymbol{Y}=y \mid \alpha,\rho,p]\\
\leqslant&B_{Y,p,\rho}\frac{1}{\alpha}\frac{\Gamma(\alpha+y_1)}{\Gamma(\alpha)}\prod_{1<j\leqslant n}\frac{\Gamma(\alpha+y_j)}{\Gamma(\alpha)}\sum_{\xi=0}^{y_j\wedge y_{j-1}}\frac{\Gamma(\alpha)}{\Gamma(\alpha+\xi)}\\
=&\frac{B_{Y,p,\rho}\Gamma(\alpha+y_1)}{\alpha\Gamma(\alpha)}\left\{\prod_{\substack{1<j\leqslant n\\ y_j=0}}\frac{\Gamma(\alpha+y_j)}{\Gamma(\alpha)}\sum_{\xi=0}^{y_j\wedge y_{j-1}}\frac{\Gamma(\alpha)}{\Gamma(\alpha+\xi)}\right\}\\
&\qquad\cdot\left\{\prod_{\substack{1<j\leqslant n\\ y_j>0}}\frac{\Gamma(\alpha+y_j)}{\Gamma(\alpha)}\sum_{\xi=0}^{y_j\wedge y_{j-1}}\frac{\Gamma(\alpha)}{\Gamma(\alpha+\xi)}\right\}\\
=&\frac{B_{Y,p,\rho}\Gamma(\alpha+y_1)}{\alpha\Gamma(\alpha)}\cdot 1\cdot\left\{\prod_{\substack{1<j\leqslant n\\ y_j>0}}\Gamma(\alpha+y_j)\sum_{\xi=0}^{y_j\wedge y_{j-1}}\frac{1}{\Gamma(\alpha+\xi)}\right\}\\
=&\frac{B_{Y,p,\rho}\Gamma(\alpha+y_1)}{\alpha\Gamma(\alpha)}\left\{\prod_{\substack{1<j\leqslant n\\ y_j>0,y_{j-1}>0}}\Gamma(\alpha+y_j)\sum_{\xi=1}^{y_j\wedge y_{j-1}}\frac{1}{\Gamma(\alpha+\xi)}\right\}\\
&\qquad\cdot\left\{\prod_{\substack{1<j\leqslant n\\ y_j>0,y_{j-1}=0}}\Gamma(\alpha+y_j)\frac{1}{\Gamma(\alpha)}\right\} \qquad (*)
\end{aligned}
$$

where $B_{Y,p,\rho}$ is a bound that depends on data Y, value of p or β and bounds of ρ.

When $\mu\to 0$ or $\alpha\to 0$, $\Gamma(\alpha)\to\infty$, $\alpha\Gamma(\alpha)\to 1$, so (*) can be bounded, and therefore $\pi(\mu,\rho \mid \boldsymbol{Y}=y)$ is bounded on the set $[0,M_\rho]\times(0,M_\mu^1]$, which implies

that the posterior is integrable.

7.

$$I_3 = \int_0^{M_\rho} \int_{M_\mu^2}^{\infty} \pi(\mu, \rho \mid \boldsymbol{Y} = y)\, d\mu d\rho$$

Similar to I_1, on the set $[0, M_\rho] \times [M_\mu^1, \infty)$, ρ and $\dfrac{\rho}{(1-\rho)^2}$ are still bounded. In addition, all terms involving only $\{y_i\}_{i=1}^n$, ρ and/or p are all bounded. Furthermore,

$$\frac{\Gamma(\alpha + y_j)}{\Gamma(\alpha)} = (\alpha + y_j - 1)(\alpha + y_j - 2) \cdots \alpha \leqslant \alpha^{y_j} \quad \text{for large } \alpha.$$

Hence,

$$\begin{aligned}
&\pi(\mu, \rho \mid \boldsymbol{Y} = y) \\
\propto &\frac{1}{\alpha} \mathrm{P}[\boldsymbol{Y} = y \mid \alpha, \rho, p] \\
\leqslant &B_{Y,p,\rho} \alpha^{B_Y} p^\alpha \prod_{1<j\leqslant n} \frac{p^\alpha}{(1-\rho+\rho p)^\alpha} \\
\leqslant &B_{Y,p,\rho} \alpha^{B_Y} r^\alpha,
\end{aligned}$$

where $r = \left(\dfrac{p}{1-\rho+\rho p}\right)^n < 1$, $B_{Y,p,\rho}$ is a bound that depends on data Y, value of p or β and bounds of ρ and B_Y is an integer that only depends on Y. When $\alpha \to \infty$, apply L'Hôpital's rule,

$$\lim_{\alpha\to\infty} \frac{\alpha^{B_Y} r^\alpha}{\frac{1}{\alpha^2}} = \lim_{\alpha\to\infty} \frac{(B_Y+2)\alpha^{B_Y+1}}{-\log(r) r^{-\alpha}} = \cdots = \lim_{\alpha\to\infty} \frac{(B_Y+2)!}{(-\log(r))^{B_Y+2}} r^\alpha = 0,$$

which means that $\alpha^{B_Y} r^\alpha$ converges to 0 faster than $\dfrac{1}{\alpha^2}$. As $\displaystyle\int_{M_\mu^1}^{\infty} \frac{1}{\alpha^2}\, d\alpha < \infty$, it is straitforward to deduce that $I_3 = \displaystyle\int_0^{M_\rho} \int_{M_\mu^2}^{\infty} \pi(\mu, \rho \mid \boldsymbol{Y} = y)\, d\mu d\rho < \infty$.

8.

$$I_5 = \int_{M_\rho}^{1} \int_{M_\mu^1}^{M_\mu^2} \pi(\mu, \rho \mid \boldsymbol{Y} = y)\, d\mu d\rho$$

By making change of variables, we can transform the $\mu\rho$-plane into $\alpha\rho$-plane:

$$\int_{M_\rho}^{1} \int_{M_\mu^1}^{M_\mu^2} \pi(\mu, \rho \mid \boldsymbol{Y} = y)\, d\mu d\rho$$

$$\propto \int_{M_\rho}^{1} \int_{M_\mu^1}^{M_\mu^2} \frac{1}{\mu} \mathrm{P}[\boldsymbol{Y} = y \mid \alpha, \rho, p]\, d\mu d\rho$$

$$= \int_{M_\rho}^{1} \int_{\beta M_\mu^1}^{\beta M_\mu^2} \frac{1}{\alpha} \mathrm{P}[\boldsymbol{Y} = y \mid \alpha, \rho, p]\, d\alpha d\rho.$$

Since α is bounded, all terms involving only α, y_j and/or p can be bounded. Additionally, $(1 - \rho + \rho p)^{\alpha + y_j + y_{j-1}}$ and ρ^ξ can also be bounded because ρ is bounded.

$$\begin{aligned}
&\frac{1}{\alpha} \mathrm{P}[\boldsymbol{Y} = y \mid \alpha, \rho, p] \\
\leqslant & B_{Y,p,\alpha,\rho} \prod_{1<j\leqslant n} (1-\rho)^{y_j + y_{j-1}} \sum_{\xi=0}^{y_j \wedge y_{j-1}} \frac{1}{(1-\rho)^{2\xi}} \\
= & B_{Y,p,\alpha,\rho} \prod_{1<j\leqslant n} (1-\rho)^{2y_j \wedge y_{j-1} + |y_j - y_{j-1}|} \cdot \frac{(1-\rho)^2 - \frac{1}{(1-\rho)^{2y_j \wedge y_{j-1}}}}{(1-\rho)^2 - 1} \\
= & B_{Y,p,\alpha,\rho} \prod_{1<j\leqslant n} \frac{(1-\rho)^{|y_j - y_{j-1}|} - (1-\rho)^{2y_j \wedge y_{j-1} + |y_j - y_{j-1}| + 2}}{1 - (1-\rho)^2}, \qquad (\#)
\end{aligned}$$

where $B_{Y,p,\alpha,\rho}$ depends on p, data Y, upper and lower limits of α and ρ.

In the product of (#), for some j, $y_j = y_{j-1}$, we define $0^0 = 1$ so that $(1 - \rho)^{|y_j - y_{j-1}|}$ is well defined for all j when $\rho \to 1$. In addition, when $\rho \to 1$, at least one terms in the product approaches 0, and thus (#) can be bounded by

a finite number when ρ is in the interval $[M_\rho, 1]$. Therefore,

$$\int_{M_\rho}^{1}\int_{\beta M_\mu^1}^{\beta M_\mu^2} \frac{1}{\alpha}\mathrm{P}[\boldsymbol{Y}=y \mid \alpha,\rho,p]\,d\alpha d\rho < \infty.$$

9.

$$\begin{aligned}
I_4 &= \int_{M_\rho}^{1}\int_{0}^{M_\mu^1} \pi(\mu,\rho \mid \boldsymbol{Y}=y)\,d\mu d\rho \\
&\propto \int_{M_\rho}^{1}\int_{0}^{M_\mu^1} \frac{1}{\mu}\mathrm{P}[\boldsymbol{Y}=y \mid \alpha,\rho,p]\,d\mu d\rho \\
&= \int_{M_\rho}^{1}\int_{0}^{\beta M_\mu^1} \frac{1}{\alpha}\mathrm{P}[\boldsymbol{Y}=y \mid \alpha,\rho,p]\,d\alpha d\rho \\
&\leqslant \int_{M_\rho}^{1}\int_{0}^{\beta M_\mu^1} B_{Y,p,\rho}\frac{\Gamma(\alpha+y_1)}{\Gamma(\alpha)} \\
&\qquad \cdot \prod_{1<j\leqslant n}\frac{\Gamma(\alpha+y_j)}{\Gamma(\alpha)}(1-\rho)^{y_j+y_{j-1}}\sum_{\xi=0}^{y_j\wedge y_{j-1}}\frac{\Gamma(\alpha)}{\Gamma(\alpha+\xi)}\frac{\rho^\xi}{(1-\rho)^{2\xi}}\,d\alpha d\rho, \quad (\dagger)
\end{aligned}$$

where $B_{Y,p,\rho}$ is a bound for y_j, p^α, $(1-p)^{y_j}$, $(1-\rho+\rho p)^{\alpha+y_j+y_{j-1}}$ and $\left(\dfrac{p^2}{1-p}\right)^\xi$.

Similar to the arguments for I_1 and I_5, the integrand of (†) can be bounded by a finite number, and thus $I_4 < \infty$.

10.

$$\begin{aligned}
I_6 &= \int_{M_\rho}^{1}\int_{M_\mu^2}^{\infty} \pi(\mu,\rho \mid \boldsymbol{Y}=y)\,d\mu d\rho \\
&\propto \int_{M_\rho}^{1}\int_{\beta M_\mu^2}^{\infty} \frac{1}{\alpha}\mathrm{P}[\boldsymbol{Y}=y \mid \alpha,\rho,p]\,d\alpha d\rho \\
&\leqslant \int_{M_\rho}^{1}\int_{\beta M_\mu^2}^{\infty} \frac{B_{Y,p,\rho}}{\alpha}\frac{\Gamma(\alpha+y_1)}{\Gamma(\alpha)}p^\alpha \\
&\qquad \cdot \prod_{1<j\leqslant n}\frac{\Gamma(\alpha+y_j)}{\Gamma(\alpha)}\frac{p^\alpha}{(1-\rho+\rho p)^\alpha}(1-p)^{y_j+y_{j-1}}\sum_{\xi=0}^{y_j\wedge y_{j-1}}\frac{1}{(1-\rho)^{2\xi}}\,d\alpha d\rho, \quad (\bigstar)
\end{aligned}$$

where $B_{Y,p,\rho}$ is a bound for y_j, $(1-p)^{y_j}$, $(1-\rho+\rho p)^{y_j+y_{j-1}}$, ρ^{ξ} and $\left(\frac{p^2}{1-p}\right)^{\xi}$. Similar to the argument for I_5, the integral with respect to ρ in ($\bigstar$) is bounded, and then following the argument for I_3, the integral with respect to α is also finite, so $I_6 < \infty$.

11. Each double integral on one of the six sets is integrable, and the integral of the posterior distribution is the sum of the six integrals over subregions, so the posterior has a finite integral, which indicates that the posterior is proper.

Bibliography

Andrews, D. W. K. (1993), "Tests for parameter instability and structural change with unknown change point," *Econometrica*, 61, 821–856.

Andrieu, C. and Thoms, J. (2008), "A tutorial on adaptive MCMC," *Statistics and Computing*, 18, 343–373.

Atchade, Y. F. and Liu, J. S. (2010), "The Wang-Landau algorithm in general state spaces: Application and convergence analysis," *Statistica Sinica*, 20, 209–233.

Bai, J. (1997), "Estimation of a change point in multiple regression models," *Review of Economics and Statistics*, 79, 551–563.

Brodsky, B. E. and Darkhovsky, B. S. (1993), *Nonparametric methods in change point problems*, Springer.

Casella, G. and George, E. I. (1992), "Explaining the Gibbs sampler," *The American Statistician*, 46, 167–174.

Chib, S. (1998), "Esmiation and comparison of multiple change-point models," *Journal of Econometrics*, 86, 221–241.

Craiu, R. V., Rosenthal, J., and Yang, C. (2009), "Learn from the neighbor: Parallel-chain and regional adaptive MCMC," *Journal of American Statistical Association*, 104, 1454–1466.

Daley, D. J. (1988), *An Introduction to the Theory of Point Processess*, Springer.

Diggle, P. J. (2003), *Statistical Analysis of Spatial Point Patterns*, Hodder Education Publishers.

Eden, U. T., Frank, L. M., Barbieri, R., Solo, V., and Brown, E. N. (2004), "Dynamic analysis of neural encoding by point process adaptive filtering," *Neural Computation*, 16, 971–998.

Edwards, C. B. and Gurland, J. (1961), "A class of distributions applicable to accidents," *Journal of the American Statistical Association*, 56, 503–517.

Engle, R. F. and Lunde, A. L. (2003), "Trades and quotes: A bivarite point process," *Journal of Financial Econometrics*, 1, 159–188.

Gatrell, A. C., Bailey, T. C., Diggle, P. J., and Rowlingson, B. S. (1996), "Spatial point pattern analysis and its application in geographical epidemiology," *Transactions of the Institute of British Geographers*, 21, 256–274.

Gelfand, A. E. and Smith, A. F. M. (1990), "Sampling-based approaches to calculating marginal densities," *Journal of American Statistical Association*, 85, 398–409.

Green, P. J. (1995), "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, 82, 711–732.

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning*, Springer.

Hastings, W. K. (1970), "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, 57, 97–109.

Ji, C. and Schmidler, S. C. (2013), "Adaptive Markov chain Monte Carlo for Bayesian variable selection," *Journal of Computational and Graphical Statistics*, 22, 708–728.

Kou, S. C., Zhou, Q., and Wong, W. H. (2006), "Equi-energy sampler with applications in statistical inference and Statistical mechanics," *Annals of Statistics*, 34, 1581–1619.

Kushner, H. J. and Yin, G. G. (1997), *Stochastic Approximation Algorithms and Applications*, Springer-Verlag.

Laplace, P. S. (1974), "Memoir on the probability of the causes of events," Mémoires de Mathématique et de Physique, Tome Sixième. (English translation by S. M. Stigler 1986. Statist. Sci., 1(19):364–378).

Law, R., Illian, J., Burslem, D. F. R. P., Gratzer, G., Gunatilleke, C. V. S., and Gunatilleke, I. A. U. N. (2009), "Ecological information from spatial patterns of plants: Insights from point process theory," *Journal of Ecology*, 97, 616–628.

Lee, H. (1998), "Model Selection and Model Averaging for Neural Networks," Ph.D. thesis, Carnegie Mellon University.

Liang, F. (2005), "A generalized Wang-Landau algorithm for Monte Carlo computation," *Journal of the American Statistical Association*, 100, 1311–1327.

MacKay, D. J. C. (1992), "A practical Bayesian framework for backpropagation networks," *Neural Computation*, 4, 448–472.

MacKay, D. J. C. (1995), "Probable networks and plausible predictions—A review of practical Bayesian methods for supervised neural networks," *Network: Computation in Neural Systems*, 6, 469–505.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953), "Equations of state calculations by fast computing machines," *Journal of Chemical Physics*, 21, 1087–1092.

Neal, R. (1997), *Bayesian Learning for Neural Networks*, Springer-Verlag.

Ogata, Y. (1999), "Seismicity analysis through point-process modeling: A review," *Pure and Applied Geophysics*, 155, 471–507.

Penttinen, A. and Stoyan, D. (2000), "Recent applications of point process methods in forestry statistics," *Statistical Science*, 15, 61–78.

Raftery, A. E. and Akman, V. E. (1986), "Bayesian analysis of a Poisson process with a change-point," *Biometrika*, 73, 85–89.

Ripley, B. (1996), *Pattern Recognition and Neural Networks*, Cambridge University Press.

Robbins, H. and Monro, S. (1951), "A stochastic approximation method," *Annals of Mathematical Statistics*, 22, 400–407.

Scargle, J. D. and Babu, G. J. (2003), "Point process in astronomy: Exciting events in the universe," *Handbook of Statistics*, 21, 795–825.

Schmidler, S. C. (2012), "A note on mixture kernels for Metroplis-Hastings chains," in preparation.

Schmidler, S. C. and Woodard, D. (2013), "Lower bounds on the convergence of adaptive MCMC methods," Under invited revision for Annals of Statistics. Originally published as an ORIE technical report in Jan 2011.

Wang, F. and Landau, D. (2001), "Efficient, multiple-range random walk algorithm to calculate the density of states," *Physical Review Letters*, 86, 2050–2053.

Wolpert, R. L. and Brown, L. D. (2011), "Markov infinitely-divisible stationary time-reversible integer-valued processes," Tech. Rep. 2011-11, Duke University Department of Statistical Science.

Young, S. R., Sparks, R. S. J., Aspinall, W. P., Lynch, L. L., Miller, A. D., Robertson, R. E. A., and Shepherd, J. B. (1998), "Overview of the eruption of Soufriere Hills Volcano, Montserrat, 18 July 1995 to December 1997," *Geophysical Research Letters*, 25, 3389–3392.

Young, T. Y. and Kuo, L. (2001), "Bayesian binary segmentation procedure for a Poisson process with multiple changepoints," *Journal of Computational and Graphical Statistics*, 10, 772–785.

Biography

Jianyu Wang was born in 1982 and grew up in Handan, Hebei province, China. She attended Nankai University and graduated with a B. S. in Mathematics in 2006. She continued her graduate study at the University of Utah and earned a Master degree in Mathematics in 2008. She joined the Department of Statistical Science of Duke University afterwards, and earned Master degree in Statistics in 2010 *enroute* to complete her Ph.D. degree.